Universidade Nova de Lisboa
Faculdade de Ciências e Tecnologia
Departamento de Informática

Dissertação de Mestrado em Engenharia Informática
2007/2008

# Extracção Automática de Informação e Conhecimento em Textos no Âmbito B2B

26376 - Júlio Flávio Tavares Barbas

Orientador
Prof. Doutor Nuno Cavalheiro Marques

Dezembro de 2008

Nº do aluno: 26376
Nome: Júlio Flávio Tavares Barbas

Título da dissertação:
Extracção Automática de Informação e Conhecimento em Textos no Âmbito B2B

Palavras-Chave:

- B2B
- Processamento de linguagem natural
- Construção Civil
- Segmentação de palavras
- Anotação morfossintáctica

Keywords:

- Business to Business
- Natural language processing
- Civil engineering
- Tokenization
- Part-of-Speech Tagging

## Resumo

Nesta dissertação é feito um estudo sobre a validade da aplicação de técnicas de processamento de língua natural a textos na área da construção civil, disponibilizados pelo portal *econstroi* gerido pela empresa Vortal. Estes textos correspondem a especificações para artigos disponibilizados por operadores deste portal. As especificações encontram-se organizadas segundo uma lista de articulado e relacionam-se com um sistema de categorização estruturado.

Numa primeira fase é feito um levantamento exaustivo sobre algumas amostras de diversos artigos, de modo a determinar a validade da utilização de ferramentas de anotação morfossintáctica no modelo de negócio do portal *econstroi*. Visto que é necessário haver um tratamento prévio dos dados para serem analisados, valida-se igualmente a utilidade de haver uma estruturação e classificação desses dados com base num sistema de categorias já existente, relacionando assim de forma sólida os dados inseridos pelos utilizadores com um sistema estruturado por categorias. É igualmente abordada a questão das ontologias, tendo por base a verificação da necessidade de relacionar estes dados textuais com uma organização de conhecimento estruturada.

Numa segunda fase, e tendo os textos anotados com etiquetas específicas e genéricas serão utilizadas gramáticas de cláusulas definidas (DCGs) para extrair conhecimento destes textos, sendo analisadas possíveis vantagens desta abordagem.

É seguida uma abordagem baseada em SOA, que possibilitará a validação das técnicas propostas e da sua interoperabilidade com os restantes módulos SOA do portal *econstroi*. Em concreto foi desenvolvido um primeiro protótipo que utiliza algumas das técnicas aqui abordadas.

## Abstract

The purpose of this dissertation is to present a study on the validity of the application of natural language processing techniques in civil Engineering Texts. These texts are available on *econstroi* portal, managed by Vortal Company. Available texts match the specifications of articles that are given by portal users. These specifications are organized according to an article list and relate themselves with a structured categorized system.

On a first phase diverse text samples are studied, as to determine the validity of using part-of-speech tagger tools on *econstroi*'s business model. The utility of having a classification and structuration based on a existent category system is also validated. Other ontologies are also studied, regarding the need to relate these textual data, with a structured knowledge organization.

On a second phase and having the texts with specific annotation tags, definite clause grammars (DCGs) will be used to extract knowledge of these texts. Possible advantages of this approach are analyzed.

A SOA approach is followed, and will permit the validation of proposed techniques, the construction of a first prototype and their interoperability with the remaining SOA modules that compound *econstroi* portal.

# Índice

## Índice de Figuras

# Índice de Tabelas

# 1. Introdução

## 1.1 Introdução geral

Na construção civil é necessário tratar grandes volumes de texto, visto ser uma área em que todos os processos operacionais e todo o material envolvido têm de estar descritos formalmente. Esta informação é muitas vezes partilhada entre diversas organizações intervenientes numa determinada obra ou construção. Inerente a esta situação está o contexto B2B (*Business to Business*), no qual se insere o portal *econstroi*, que nos servirá como base de estudo. Este portal faz parte da plataforma Vortal[1], e funciona como um mercado electrónico (sobre Internet) dirigido às empresas do sector da Construção, para fazerem negócios (comércio electrónico) de uma forma mais rápida, simples e eficaz.

No portal *econstroi* é possível criar textos contendo listas de artigos: produtos para obras e projectos da construção civil, que são enviadas às empresas fornecedoras, de modo a que estas possam analisar os articulados e fornecer esses produtos. Estas listas de artigos detalham as características dos produtos e das suas propriedades, tendo estes artigos uma categoria que os caracteriza, com base num sistema de categorização da Vortal. É apresentado um exemplo de uma descrição destes artigos na secção 2.3.1.

Um dos problemas deste tipo de textos é a inexistência de uma terminologia consistente que possa ser utilizada globalmente, enquanto protocolo, entre as diversas organizações que necessitam de comunicar entre si de uma forma textual. Há uma grande pluralidade de classificações de elementos e componentes no ramo da construção civil, tornando os processos muito mais complicados. Desta forma vão sendo criadas perdas

[1] Vortal, S.A. – Empresa que gere um portal B2B do ramo da Construção Civil. (www.vortal.biz)

qualitativas em todos os processos sempre que há comunicação entre os diversos intervenientes, podendo haver um decréscimo de produtividade e qualidade no resultado final da obra ou produto.

Visto que é necessário haver um tratamento prévio dos dados para serem analisados, estudou-se a utilidade de haver uma estruturação e classificação com base num sistema de categorização de produtos, relacionando assim de forma sólida os dados inseridos pelos utilizadores com um sistema estruturado por categorias e/ou ontologias.

É seguida uma abordagem baseada no processamento superficial do Português [30], utilizando etiquetadores morfossintácticos [14, 16] aplicados aos textos das descrições de artigos. É feito um levantamento sobre a validade da utilização de ferramentas de anotação morfossintáctica, tendo como base um subconjunto com uma amostra de texto, contendo descrições de artigos. Para tal, estas ferramentas foram integradas seguindo um modelo SOA (*Service Oriented Architecture*), segundo o qual o portal está projectado.

## 1.2 Descrição e contexto do problema em estudo

O problema em estudo tem por base a aplicação de um sistema de anotação morfossintáctica [1, 2], auxiliado por conhecimento do domínio [14] da área de negócio da Vortal.

Em concreto verifica-se a utilidade da utilização de etiquetas genéricas e específicas na classificação dos textos, para posterior tratamento e extracção dos dados mais relevantes. Para isso será efectuado um estudo sobre amostras significativas de diversos textos, de modo a determinar a validade da utilização destas ferramentas de anotação morfossintáctica no modelo de negócio da Vortal.

Uma vez os textos anotados, são utilizadas gramáticas de cláusulas definidas (*Definite Clause Grammar* - DCG), para expressar relações gramaticais específicas para o problema geral em estudo. Desta forma possibilita-se a extracção de conhecimento destes textos, com base em regras que definem estruturas de frases. Estas regras têm o objectivo de extrair informação estruturada com base nos textos anotados. Assim, são analisadas as possíveis vantagens desta abordagem, no que respeita à utilização de informação dos

textos anotados. A utilização destes textos possibilita a obtenção ou a análise, através das DCGs, de novas frases, que seguem uma estrutura previamente definida, mas com palavras e expressões da base de conhecimento em estudo. Da mesma forma, é possível utilizar estas regras gramaticais para validar frases e expressões.

Com base nestes textos, cuja informação foi extraída da base de dados e posteriormente analisada e anotada, foi desenhado um modelo que permite a criação de um *Data Warehouse*.

Esta integração entre a plataforma Vortal e os serviços utilizados na análise morfossintáctica é possível segundo um modelo SOA. Este modelo permite a fácil consulta dos dados da plataforma e a fácil transformação dos mesmos para tratamento futuro. Em concreto foi desenvolvido um primeiro protótipo que utiliza algumas das técnicas aqui abordadas. Igualmente torna-se possível e simples efectuar testes, bastando integrar serviços de teste neste modelo, de modo a que comunique tanto com os dados da plataforma como com os serviços de análise textual.

O estudo assenta sobretudo em conceitos de *Text Mining*, tendo como finalidade extrair informações relevantes de uma grande base de textos, sem a necessidade de os ler previamente, auxiliando na navegação para encontrar a informação pretendida. A anotação morfossintáctica é feita com base num etiquetador de texto [3,4], possibilitando a classificação de expressões e a associação de palavras no texto a itens da base de conhecimento.

Visto que a área de negócio em causa tem particularidades exclusivas no que respeita ao tipo de textos e expressões, a extracção de conhecimento, com base em DCGs [24], tem de ser feita tendo em conta esse factor e no objectivo final, estruturando o resultado de forma válida para o utilizador.

Para utilizar todas estas ferramentas foi ainda criada uma representação contendo a informação obtida nos diversos módulos, necessários para efectuar a extracção da informação no texto. Nomeadamente, a normalização dos dados, a segmentação de palavras, a anotação morfossintáctica, a anotação por categorias específicas da área de negócio em estudo, o armazenamento da informação classificada e anotada e a extracção de conhecimento útil e inteligente para o utilizador final [6, 11]. Assim, verificamos que

o problema em estudo assenta num processo completo de anotação, extracção de informação e sua relação com uma base de conhecimento, tendo como foco a aplicabilidade destas ferramentas num determinado contexto específico de negócio [16,17].

Todas estas ideias e conceitos genéricos unem-se para formar uma solução específica, cuja necessidade surge do fraco desenvolvimento actual de soluções relacionadas com a análise e extracção de informação relevante, para o utilizador final, na actual plataforma Web da Vortal [4]. Desta forma, potencia-se a utilização de ontologias [7,9,18]. Possibilita-se ainda a fácil geração de novo conhecimento a partir da imposição de normas fixas e geradas dinamicamente no portal. O objectivo desta integração é o de fornecer ao utilizador pesquisas mais eficientes, ajuda automática na escrita de documentos técnicos [29] e construir modelos de dados que podem gerar novo conhecimento com base nesses dados estruturados. Assim, esta tese foi motivada pelo seguintes pontos:

1. Identificar um conjunto de vantagens das técnicas de extracção de informação do texto numa abordagem B2B;
2. Identificar casos de estudo em que a utilização de técnicas de análise sintáctica contribui para a extracção de informação do texto;
3. Validar a utilidade da adaptação de técnicas gerais de análise do português com informação específica da área da construção;
4. Aplicação de ferramentas de *Business Intelligence* à base de conhecimento Vortal, incluindo informação que foi extraída do texto.

É com base nestes aspectos que ao longo deste estudo são salientadas técnicas de extracção de informação, cuja aplicação se enquadra essencialmente num modelo B2B. Isto porque se lida num contexto de negócio, onde existem diversos intervenientes que alimentam o sistema com um fluxo de partilha de dados. É tendo em conta estes aspectos que são identificados casos de estudo, onde é possível proceder à análise dos dados e consequente extracção de informação. Estes casos de estudo são principalmente as fases

do processo comunicativo entre os intervenientes, que mais se adequam ao processamento da informação.

Neste sentido, torna-se possível aplicar técnicas, que são gerais de análise do português, aos dados do portal. Visto que estes dados são constituídos por informação textual relativa à área da construção civil, é importante notar o seu carácter específico, enquanto sub-linguagem, contendo determinados termos e construções sintáctico-semânticas específicas. Deste modo, é importante a adaptação que se faz com base na utilização de etiquetas específicas ao tema da construção, para que se possa fazer uma análise mais refinada e específica.

Ainda neste contexto relativo à análise da informação, surge o conceito de *Business Intelligence* (BI), e as questões relativas à integração deste tipo de ferramentas com a extracção de conhecimento feita através do *Text Mining* [31]. Nesta fase é proposto um processo que permite alimentar uma base de dados de BI (ou seja, um *Data Warehouse*) com a informação estruturada obtida através da análise textual. Como resultado, obtém-se a possibilidade de analisar e consultar a informação, gerando assim estatísticas e novo conhecimento relacionado com esses dados extraídos.

Tendo em conta o processo descrito anteriormente, importa enquadrar o que se entende por conhecimento e por informação. No processo inicial é feita uma extracção de informação, visto que é um processo que visa obter dados, estruturá-los e anotá-los. Tendo os dados correctamente anotados com etiquetas específicas e genéricas, é possível obter novos dados, que à partida não existem, com base em gramáticas. Assim este processo pode ser considerado como extracção de informação. No caso do modelo de *Data Warehouse*, as *querys* que se fazem serão consideradas extracção de conhecimento, sempre que sejam imediatamente úteis para os processos de decisão da empresa.

## 1.3 Solução apresentada

Visto que o etiquetador de texto está implementado em Prolog houve a necessidade de criar uma interface que pudesse comunicar de uma forma directa e flexível com os

programas implementados em Prolog e que estivesse inserida na plataforma .NET 2.0, de modo a tornar o desenvolvimento de programas a nível da camada de aplicabilidade mais fácil e intuitivo.

Para isso, foram feitas diversas pesquisas sobre possíveis ferramentas que permitissem abstrair a interacção referida. Existem duas tentativas de abordar este problema, uma implementação de raiz do Prolog em C#: P#[2] e outra que utiliza algumas bibliotecas de comunicação com o Prolog utilizando código em C: CSharpInterface[3]. Embora interessantes estas abordagens, nenhuma delas é exactamente o que se pretendia. Assim foi projectada e criada uma classe C# que expõem métodos para comunicar com um simples processo Prolog.

Esta classe pode ser utilizada em qualquer programa C#, seja *Windows Forms*, *ASPX* ou *Web Service*, permitindo uma fácil adaptação a grande parte dos ambientes existentes hoje em dia.

As diversas ferramentas de anotação morfossintáctica já implementadas em Prolog [1,3,4] podem, desta forma, ser facilmente utilizadas por outros programas implementados em C#, bastando para isso usar esta classe que abstrai a funcionalidade da linha de comandos do SWI-Prolog.

Para ser possível utilizar este sistema de anotação, e relembrando que os dados textuais disponíveis estão em bruto na base de dados do sistema actual, foi necessário tratar e normalizar esses dados antes de os submeter ao etiquetador morfossintáctico. Os textos são previamente separados por frases, para depois serem separadas em palavras que estão prontas para serem analisadas e anotadas. Esta anotação é feita empregando um etiquetador morfossintáctico utilizando redes neuronais, previamente treinadas, e dicionários [17] adaptados ao domínio da construção civil, como descrito anteriormente na secção 1.1 deste capítulo.

Tendo os dados estruturados e analisados com o auxílio de DCGs, é possível extrair conhecimento lógico para disponibilizar ao utilizador final. Na Figura 1.1 ilustra-se os processos descritos anteriormente. Especificamente, e tendo em conta os dados já analisados, é possível relacioná-los com bases de conhecimento (por vezes representados

[2] *P#: A concurrent Prolog for the .NET framework* (http://www.dcs.ed.ac.uk/home/stg/Psharp/)
[3] *CSharp Interface to SWI-Prolog* (http://gollem.swi.psy.uva.nl/twiki/pl/bin/view/Foreign/CSharpInterface)

como ontologias). Ou seja, são criadas estruturas de dados bem definidas e inter-relacionadas por meio de conceitos, obtendo assim dados para alimentar esta base de conhecimento. No estudo aqui efectuado foi modelado um pequeno *Data Warehouse*, potenciando a utilização de ferramentas de *Business Intelligence* para extrair todo o tipo de informação estatística e organizacional.

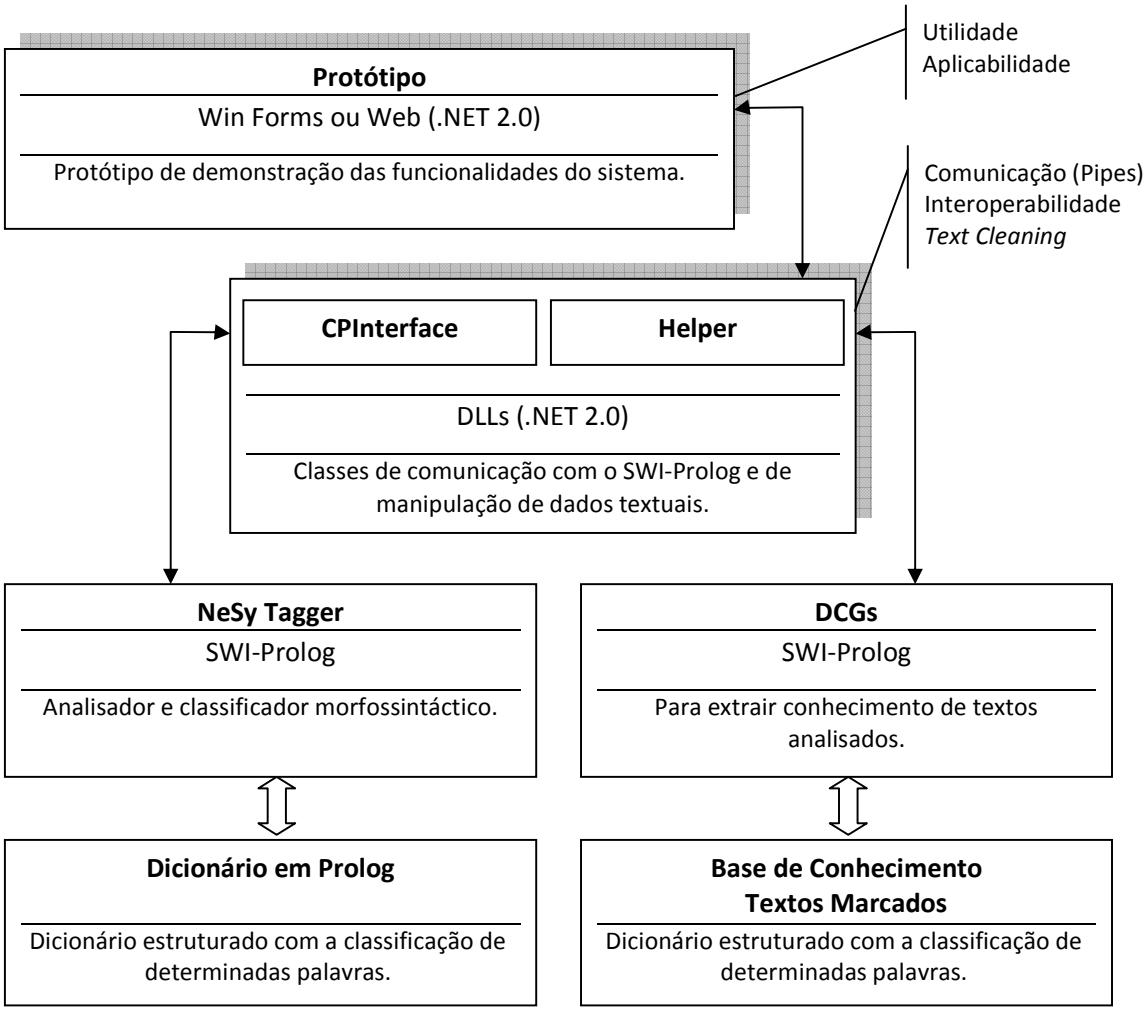



Figura 1.1 – Figura da solução apresentada.

## 1.4 Principais contribuições

Uma das principais contribuições é a proposta de um modelo, e demonstração da sua utilidade, para tratamento de texto, tendo em conta o tipo de negócio da Vortal. Este modelo tem por base a extracção automática de informação e conhecimento em textos, por relação com uma base de conhecimento já existente no portal *econstroi*.

Tendo os dados classificados e anotados com etiquetas específicas, e com o auxílio de gramáticas (DCGs) específicas para o problema em causa, obtém-se: sugestões de frases e expressões para uma determinada descrição, templates relativos a medidas e referências de diversos materiais constituintes na descrição do artigo. Desta forma, possibilita-se uma fácil implementação de novas funcionalidades tendo como base texto já anotado. Assim, como primeiro objectivo desta tese:

**Objectivo 1:** Estudar de que forma a criação das etiquetas específicas contribui para a melhor análise aos textos, no que respeita à capacidade de extrair e validar informação específica da área da construção civil (Capítulo 4).

Com base nas DCGs e nos textos anotados são criados módulos que permitem a extracção de novo conhecimento a partir destes textos. Estes novos módulos podem ser integrados em módulos actuais do portal *econstroi*. São desenvolvidos mecanismos que permitem interligar e integrar tecnologias diferentes seguindo o conceito SOA (*Service-oriented Architecture*), possibilitando a incorporação de ferramentas em estudo com funcionalidades específicas no portal B2B *econstroi*, relativas a análise e interligação de informação dispersa em textos. Assim surge o segundo objectivo desta tese:

**Objectivo 2:** Estudar a aplicabilidade destas ferramentas a este e a novos domínios, nomeadamente no domínio do *Busines Intelligente*, e na criação de modelos que permitem a análise dos dados extraídos de uma forma relacional. Em concreto, é construído um *Data Warehouse* que conjuga o conhecimento na base de dados Vortal com informação extraída do texto. Um primeiro protótipo é apresentado na secção 3.7.

A SOA possibilita o desenvolvimento imediato de um conjunto de módulos que são simultaneamente úteis para o negócio da Vortal e para o estudo da aplicabilidade realizado nesta tese, permitindo também a, eventual, e futura fácil integração do protótipo proposto no modelo de negócio da Vortal. Desta forma, surge também um terceiro objectivo:

**Objectivo 3:** Utilizar serviços SOA para construção de várias aplicações e serviços que utilizem conhecimento derivado do texto. Nomeadamente, foi construído um protótipo que auxilia o preenchimento de novas descrições, com base em itens da base de conhecimento da Vortal.

A utilização do SOA possibilitou assim a criação de várias aplicações e serviços que podem ser directamente utilizados no âmbito da empresa. Em resumo, foram construídos protótipos para:

- Modelo relacional estruturado com os dados relativos às descrições dos artigos;
- Geração de novas descrições de artigos, com base na utilização de gramáticas e de textos anotados;
- Validação de frases relativas a descrições de artigos, com base nas diversas gramáticas definidas;
- Modelo multidimensional das transacções dos artigos, relacionados com informação temporal, com a categoria, com as unidades de medida e com os produtos.

## 1.5 Plano de leitura

No capítulo 2, é apresentado trabalho relacionado, onde se procura mostrar alguns trabalhos relevantes para o estudo desenvolvido na tese, nomeadamente seguindo os desenvolvimentos num projecto que visa a estruturação dos dados no domínio da construção civil. Neste capítulo, são também abordadas técnicas e metodologias que

servem de suporte ao estudo desta tese. Foca-se o tema da aplicabilidade, e da sua importância ao nível das arquitecturas orientadas aos serviços e do conceito B2B.

No capítulo 3, são explicados em detalhe o tipo de dados textuais utilizados nas análises e a forma como foram estruturados esses dados, para posterior anotação. É explicado de que forma foi classificado o texto e justificadas as razões que levaram a escolher as etiquetas específicas que são utilizadas. É apresentado um protótipo que faz parte da classificação das palavras, gera e reconhece novas frases com base em gramáticas e alimenta um modelo de *Data Warehouse* com base nos textos classificados.

O capítulo 4 contém a descrição e análise dos ensaios experimentais realizados, nomeadamente é focada a qualidade dos dados extraídos do texto. Estes ensaios são relativos à ambiguidade na classificação com etiquetas específicas, à análise da geração de frases e à análise de reconhecimento de frases e expressões.

No capítulo 5 é feita uma análise sobre os temas abordados nesta tese, dando especial atenção aos resultados obtidos nos ensaios experimentais. Nesta abordagem é tido em conta a metodologia SOA e as questões da ontologia, numa vertente relacionada com as etiquetas específicas, e como estas, juntamente com as gramáticas, servem para criar estruturas bem definidas de conhecimento.

# 2. Trabalho relacionado

O trabalho relacionado está dividido em três vertentes: uma vertente relativa a ontologias que vão de encontro com o problema da estruturação dos dados; uma vertente relacionada com os conceitos de *Text Mining* e análise de textos, com especial ênfase na Construção Civil; e por último uma vertente relacionada com o estudo dos conceitos relevantes a nível da aplicabilidade.

## 2.1 Ontologias no ramo da Construção Civil

Segundo [8], ontologias são estruturas de representação do conhecimento apropriadas para a organização de informação previamente não estruturada, que servem principalmente para facilitar a partilha de conhecimento e promover a sua reutilização.

Com base numa ontologia podemos estruturar a informação importante e associar-lhe determinados conceitos e símbolos que caracterizam essa informação. Desta forma, podemos posteriormente ter um mapa conceptual de toda a nossa informação, promovendo a interligação e estruturação de todos os dados disponíveis no sistema e impulsionar novas pesquisas mais complexas.

Desde que o conceito de ontologia surgiu tem havido diversos projectos que tentam criar modelos para a informação em diversas áreas do conhecimento humano. Na Construção Civil existem muitas normas ISO[4] (*International Standards Organization*) que tentam propor diversas nomenclaturas internacionais para as referências e caracterização dos artigos de construção.

[4] Ver http://www.iso.org

A indústria da construção ainda demonstra elevada carência na organização das informações de forma conjuntural e macro-estruturada [32]. Com o desenvolvimento da informática têm surgido inúmeras ferramentas que tentam ajudar o sector da construção a organizar a informação relacionada com tudo o que está envolvido nos processos de construção, desde os recursos humanos até aos projectos de obra e gestão de materiais.

Diversas ontologias têm sido propostas [10,13,14] tentando criar estruturas de dados que possam ser aceites globalmente e que sirvam os interesses de todos os intervenientes na área da construção. As características destas ontologias são explicadas em mais detalhe na secção 2.1.1 deste capítulo.

Existem vários trabalhos que demonstram este tipo de esforços, na implementação de normas que possam ser seguidas pela generalidade das empresas. De seguida será abordado um caso concreto que se considerou exemplificativo da implementação de ontologias para a construção civil. Visto que a implementação de uma ontologia requer uma estruturação complexa e bem definida, analisam-se ainda alguns trabalhos que tentam ajudar na sua criação e manutenção. Foram igualmente analisados outros sistemas semelhantes (por exemplo [14,18]) mas que não foram considerados tão relevantes para o problema em estudo nesta proposta.

### 2.1.1 Definição de ontologias

O projecto CDCON: Classificação e terminologia para a construção [2] é um sistema que tenta contribuir para a criação de uma terminologia consolidada e um processo de classificação genérico e geral para o mercado da construção. Para isso propõe uma estrutura básica para a descrição de todos os objectos do universo da construção, apresentando termos e conceitos e as suas inter-relações lógicas. Neste estudo é referida a enorme variedade de materiais, serviços e equipamentos que são provenientes de diversos sectores, tendo cada um deles características diferentes e linguagens técnicas específicas. Alguns desses sectores já têm bem definido as suas terminologias e os seus próprios códigos de referência para os seus produtos, o que acaba por gerar uma grande

dificuldade na sua caracterização e comparação a nível global, ou seja, deixa de existir uma coerência entre a nomenclatura utilizada pelas várias entidades. Um exemplo disto é a adopção de diferentes normas ISO por parte das empresas da construção civil. Neste trabalho, são feitas comparações entre ontologias e as normas rígidas ISO, sendo apresentado um exemplo da pouca versatilidade das ISO: a norma ISO TR 14177. Esta norma é específica para o tratamento de informações, sendo composta por cinco árvores classificatórias: *Construções*; *Elementos e Componentes*; *Materiais*; *Processos*; e *Atributos*. Porém, nestas árvores classificatórias existe uma sobreposição dos descritivos dos objectos. *Materiais* estão presentes tanto em produtos como nos elementos arquitectónicos, visto que é comum falar-se em casa de pedra ou prédio em aço.

Existem esforços de algumas entidades e organizações para a globalização dessas terminologias. Actualmente, ainda nenhuma norma específica conseguiu alcançar a completa aceitação no sector da construção.

Sem estes conceitos, o desenvolvimento de novas ferramentas informáticas, no âmbito da construção civil, torna-se bastante difícil. Isto porque envolve um grande esforço para conseguir englobar todas as variantes das nomenclaturas existentes, caso que poderia ser evitado se houvesse uma referência pública e universal bem estruturada. A maior dificuldade é possibilitar a comunicação entre todas as entidades envolvidas num processo de construção, já que entre elas podem ser utilizadas terminologias distintas, para os processos e produtos da construção.

Este projecto CDCON [2] teve assim como objectivo o desenvolvimento de uma terminologia e um sistema de classificação que resolva este tipo de lacunas, dando suporte ao desenvolvimento de sistemas de gestão relacionados com a informação específica da construção, com base em terminologias globalmente aceites e inter-relacionadas com produtos e serviços. Um dos focos principais foi a normalização de termos e conceitos específicos da área, criando uma ontologia que permitisse a interoperabilidade de sistemas. Verificou-se que a utilização de teorias específicas para indexação e pesquisa de documentos não serviriam todos os objectivos, pois, apesar de sistematizarem a área do conhecimento, descrevem os termos com base nas características linguísticas, e não nas conceituais. Assim, no projecto CDCON, foi

estabelecida uma base teórica para a definição de classes e facetas [2] para classificação e contribuindo para a criação de um léxico específico e respectiva ontologia.

Relativamente ao trabalho desta tese de mestrado este projecto é importante na medida em que revela e justifica a necessidade de se ter uma abordagem assente em categorias bem estruturadas, como é o caso das existentes no portal *econstroi*.

### 2.1.2 Construção de ontologias

A aplicação prática que pareceu mais relevante de todas as referências estudadas e relacionadas com o projecto CDCON é o programa ONTOARQ [27]. Este programa implementa um sistema relacional, baseando-se na estrutura geral para organizar os dados relativos à construção, proposto pelo projecto CDCON. O ONTOARQ permite a gestão de um conjunto de termos, conceitos, relacionamentos e associações. É também fornecida uma interface que permite a visualização gráfica dos dados estruturados segundo a ontologia que foi implementada.

A interface do programa ONTOARQ permite ainda a inserção de termos seguindo um determinado fluxo: o sistema processa previamente cada inserção verificando se o termo a ser inserido já existe ou se existe algum similar. Para tal, são utilizadas ferramentas de lematização (semelhantes às descritas em [30]) que possibilitam uma comparação mais ampla. Neste processo é associado o termo a determinadas categorias tendo como base um conceito específico.

Uma funcionalidade muito interessante desta implementação é a possibilidade de visualizar a rede de relacionamentos do termo que é pesquisado, as suas definições, a classificação por categorias e tipos de relacionamentos. Pode-se visualizar os nós por meio de uma representação gráfica de uma rede, que representa os termos associados à pesquisa, o seu significado e as ligações entre eles, que representam o tipo de relacionamento.

Como conclusão, este estudo fala-nos da melhoria da integração e da interoperabilidade, com vista a uma maior produtividade e qualidade no sector resultante da padronização de terminologias e da elaboração de ontologias aceites num domínio

específico do conhecimento. De facto, este programa oferece uma base comum de vocabulários e relacionamentos lógicos para o desenvolvimento de aplicações relacionadas a linguagens de textos, como catálogos de produtos ou sistemas de gestão de processos, isto porque uma ontologia deve ser facilmente acedida e manipulada por diferentes utilizadores e programas aplicacionais, de modo a enriquecer o conhecimento base e aperfeiçoar os conceitos inerentes. O ONTOARQ auxilia na compreensão e na fácil gestão dos relacionamentos entre os termos e os seus respectivos conceitos, sendo de grande utilidade na criação de ontologias e como forma de ensino e formação de técnicos do sector da construção devido à sua interface gráfica.

### 2.1.3 Utilização de ontologias

O Sistema de Indexação e Recuperação de Informação em Construção baseado numa Ontologia [1] visa melhorar a precisão na pesquisa de documentos textuais da área da construção, baseando-se na classificação de objectos do universo construído, desenvolvido no âmbito do projecto CDCON e na ontologia destes objectos proposta pelo projecto ONTOARQ, promovendo desta forma uma indexação automática da base de documentos.

Este trabalho tem como motivação a elevada diversidade de tipos de informação que está disponível hoje em dia na Internet, onde os sistemas de recuperação de informação (IRS – *Information Retrieval Systems*) têm um papel fundamental, lidando sobretudo com o armazenamento, recuperação e gestão de informações. É-nos referida a grande quantidade de estudos efectuados utilizando bases de texto extensas. Na maioria das aplicações os resultados destes sistemas estão ainda aquém dos desejados. Estes problemas surgem maioritariamente em sistemas de pesquisa genéricos, onde é muito difícil inferir sobre o contexto das palavras-chave a serem pesquisadas, visto que não consideram um contexto específico inerente a essa pesquisa.

Em [1], argumenta-se que nos sistemas de recuperação de informações textuais o vocabulário deve ser controlado, caracterizado por um conjunto finito de termos que se encontre organizado de uma forma estruturada, de modo a permitir controlar sinónimos

que indiquem relações entre os termos. Desta forma, no sistema descrito em [1], são propostas duas bases de dados distintas, uma para armazenar os documentos originais que contêm a informação relevante e a outra contém as entradas que representam os documentos do sistema. Tendo esta estrutura, o processo automático de indexação compara e identifica os termos relevantes nos documentos, inserindo-os posteriormente numa estrutura indexada. Este tipo de processos implica de resto algumas das etapas do processamento superficial do texto (p.ex. referidas em [30]): identificação de termos (simples ou compostos), normalização morfológica (lematização e *stemming*) e a selecção de termos. A esta acrescenta-se ainda uma tarefa típica nos sistemas de recuperação de informação: a remoção de *stopwords* (palavras irrelevantes na pesquisa).

O desempenho deste tipo de sistemas depende sobretudo da organização e estrutura da base de dados de referência, pois esta deve reflectir o contexto do universo a ser pesquisado, garantindo melhor precisão do sistema. É neste campo que é utilizada a estrutura para a descrição de objectos do projecto CDCON [2] e a base de dados relacional do ONTOARQ [27], a qual é estruturada por facetas, e serve como referência no processo de indexação, propiciando uma melhoria de desempenho conforme revela o estudo.

O estudo presente em [1] tem como objectivo o desenvolvimento de um sistema de recuperação de informação para o site INFOHAB. Este portal gere uma base de documentos técnicos relacionados com a área da construção civil, baseando-se na classificação de informações proposto pelo CDCON. É implementado um sistema de indexação automático para os documentos catalogados INFOHAB. No caso das pesquisas, estas passam por um processamento que identifica os termos relevantes na *query* de pesquisa, a remoção de *stopwords* e na normalização e padronização do vocabulário com base na base de dados de conhecimento do CDCON e do INFOHAB. É igualmente feita a extracção dos termos da base de dados do INFOHAB com avaliação da sua relevância e classificação dos documentos do INFOHAB usando esses termos. Neste estudo é verificado que o sistema proposto melhora a eficácia das pesquisas no âmbito da base de dados do INFOHAB. Segundo os resultados apresentados, neste caso, podemos verificar que este tipo de ferramentas de processamento de linguagem natural pode ser

aplicado em motores de busca de documentos específicos aumentando muito a eficácia das pesquisas tendo em conta a análise do contexto das palavras a serem pesquisadas e tendo como suporte uma base de conhecimento bem estruturada, indexada e classificada segundo uma ontologia específica à área em questão.

Resumindo, o estudo descrito na secção 2.1.1, define uma base teórica sólida para a implementação de uma ontologia para a construção civil. Mas, importa reflectir até que ponto se deve optar pelo esforço que é necessário para a criação de uma ontologia completa e robusta, face à maior simplicidade da criação de normas rígidas (como as normas ISO). Pode haver casos em que, apesar do maior esforço exigido na criação de uma ontologia, é obtida uma boa estrutura e contextualização dos dados, possibilitando uma melhor obtenção de informações relevantes. Noutros casos, em que os dados não têm uma importância tão grande, é possível que uma estruturação rígida tenha um desempenho aceitável para quando não se pretende obter uma extracção de conhecimento muito complexa. O estudo [2] é igualmente importante para tentar perceber que passos são necessários, para num trabalho futuro, construir uma ontologia tendo em mente o contexto do negócio B2B, e a informação que é trocada entre as entidades.

## 2.2 *Text Mining* e Anotação Morfossintáctica

O *Text Mining* refere-se ao processo de obtenção de informação de qualidade a partir de texto escrito numa linguagem natural. É inspirado no *Data Mining*, que consiste na extracção de informação de bases de dados estruturadas, sendo que o *Text Mining* extrai informação de dados não estruturados ou semi-estruturados [11].

Este processo tem ganho muita importância com o crescimento da Internet e dos mecanismos de busca, visto que se pode extrair informação relevante de uma grande quantidade de textos, sem a necessidade de os ler previamente. Na generalidade, um processo de *Text Mining* consiste na recolha de informação textual, numa etapa de pré-processamento, num mecanismo de indexação, na aplicação de um algoritmo baseado em técnicas de aprendizagem automática (p.ex. árvores de decisão ou redes neuronais), que

extrai dos dados conhecimento na forma de hipóteses e regras ou de modelos de classificação, e finalmente na análise dos resultados [11].

Neste trabalho é utilizado e estudado o sistema descrito em [16], utilizando etiquetadores morfossintácticos. Estes sistemas têm como objectivo a determinação, de uma forma não ambígua, da categoria sintáctica de cada palavra pertencente a um texto, tendo eventualmente de ter em conta o contexto da palavra. A base do etiquetador em estudo assenta sobre uma rede neuronal, havendo a necessidade desta ser previamente treinada de modo a ficar especialista na classificação dos tipos de elementos que o etiquetador pretende classificar [4, 19].

Todos os processos de processamento de linguagem natural envolvem determinados procedimentos que têm de ser levados a cabo para que se consiga obter uma boa análise dos mesmos. Em [30] é proposto um conjunto de procedimentos que nos permite obter a referida análise final a um determinado texto. São sugeridas cinco tarefas: Segmentação de Frases (*Sentence Segmentation*), Segmentação de Palavras (*Tokenization*), Anotação Morfossintáctica (*Part-of-Speech Tagging*), Traçamento Nominal (*Nominal Featurization*) e Lematização Nominal (*Nominal Lemmatization*). O autor seguiu uma abordagem baseada em processamento superficial, "segundo a qual a informação linguística é associada ao texto com base em informação local". Em [30], o que importa reter são as técnicas estudadas que servem para fazer um processamento do texto em bruto para frases mais simples e posteriormente para palavras, aplicando, finalmente, a Anotação Morfossintáctica (*Part-of-Speech Tagging*). Devido à utilização desta abordagem neste trabalho (detalhada no capítulo 3), apresentam-se de seguida as principais características do estudo realizado em [30].

Inicialmente é descrito um levantamento exaustivo dos diversos tipos de ambiguidades relacionadas com a segmentação de frases, com segmentação de palavras e com a anotação morfossintáctica que podem ocorrer num texto, tendo como foco principal o caso dos textos escritos em português. Este levantamento é de estrema importância para esta tese, visto que define previamente muitas das regas que foram adoptadas (secção 3.1 e 3.2) para separação de frases numa primeira fase do processo de tratamento dos textos.

Existem muitas ambiguidades que podem surgir relacionadas com a detecção de frases num texto, ficando a dever-se este facto às diversas utilizações dos símbolos que geralmente são utilizados como referência indicativa da terminação da frase. No caso da separação das palavras de uma frase, existem os problemas relacionados com a definição de uma palavra, visto que por vezes ocorrem situações em que uma palavra pode englobar determinados símbolos que são usados como divisores de palavras, e neste caso importa definir previamente este tipo de situações.

São também exploradas as técnicas de decomposição de determinadas palavras compostas na sua forma simples, utilizando técnicas de lematização. Um aspecto importante é a capacidade de preservar a informação, visto que muitas das técnicas utilizadas para dividir as frases e as palavras podem fazer com que se perca informação original e o sentido da frase ou palavra. Se isto acontecer torna-se impossível reverter o processo e saber qual o sentido e a forma original da frase ou palavra. Estes casos têm de ser tomados em consideração dependendo do resultado final da classificação que se pretenda atingir.

Numa perspectiva mais aplicacional, a anotação morfossintáctica já foi aplicada à construção civil. Em [29] é estudado um sistema Web de gestão, relacionado com a construção civil que utiliza anotação morfossintáctica para facilitar a pesquisa de informação (restringindo as buscas a nomes e entidades nominais). Neste sistema o processo de pesquisa inteligente de documentos é suportado por bases de conhecimento, documentos indexados, arquivo de regras, consulta original do utilizador previamente processada e os parâmetros da pesquisa. As bases de conhecimento servem para optimizar a pesquisa dos documentos, sendo compostas por sistemas de classificação *thesaurus*[5] e padronizações. Estas informações são armazenadas em estruturas de dados estruturadas, permitindo a associação às palavras-chave indexadas, agregando assim "inteligência" às pesquisas. Para o processamento das pesquisas e da indexação automática já foram criadas diversas operações sobre termos e consultas, como é o caso da Análise Lexical, Algoritmos de Pesquisa em Cadeias e Operações de Lematização. Na

[5] *Thesaurus* é uma lista indexada de palavras com significados semelhantes, dentro de um domínio específico de conhecimento.

Análise Léxica há uma transformação de uma sequência de caracteres numa sequência de palavras, servindo, sobretudo, para diminuir a quantidade de palavras com baixo potencial para a pesquisa. Os Algoritmos de Pesquisa em Cadeias de caracteres permitem encontrar padrões conhecidos no texto. As Operações de Lematização servem para diminuir a quantidade de termos a indexar, pois eliminam os sufixos, os prefixos e os plurais de cada palavra.

No sistema [29] são novamente levantadas questões relativas à representação do conhecimento de um certo domínio através de uma ontologia, referindo o problema de não existirem actualmente ontologias consolidadas na área da construção civil, mas destacando os esforços do projecto CDCON [2].

### 2.2.1 Recursos para análise da língua natural

Nos últimos anos têm-se multiplicado os trabalhos para análise de textos em língua natural. Muitos trabalhos têm sido apresentados na área da linguística computacional. Nomeadamente, refiram-se, a título de exemplo, sistemas de auxílio à indução de gramáticas (p.ex. [12], [18]) ou sistemas para análise textual e extracção de informação baseada em ontologias para *business intelligence* (p.ex. [6]).

No âmbito desta tese são particularmente relevantes os trabalhos para processamento computacional do português. Em particular os recursos disponibilizados em domínio público pelo projecto Linguateca [25]. Em particular, será utilizado o corpus anotado do CETEMPublico [28], essencial para a parametrização do etiquetador morfossintáctico [14] que serviu de base à anotação efectuada aos textos aqui analisados. Por simplicidade optou-se por construir manualmente uma gramática exemplificativa de alguns casos de estudo. Esta gramática utiliza o formalismo DCGs em Prolog [24].

O sistema GATE é um ambiente de desenvolvimento visual integrado, que suporta a execução e análise de sistemas modulares de processamento de língua natural. Ou seja, permite a criação de processos que manipulam e tratam textos. Atendendo aos objectivos deste trabalho, e à complexidade do sistema GATE, não se optou por utilizar esta

ferramenta: actualmente esta ferramenta ainda está muito direccionada para os textos em inglês, e o formato para representação do texto analisado apresentou algumas dificuldades a nível de tratamento futuro. Assim, foi criada uma ferramenta específica para tratar as amostras de texto aqui estudadas (secção 3). Espera-se que futuramente possam ser desenvolvidos novos módulos para o GATE que permitam uma maior integração com outros formatos para gravar o texto, considerando-se que esta integração será um próximo passo, que deverá ser tido em conta, na evolução do sistema aqui proposto.

Uma das grandes vantagens da utilização de formatos extensíveis para o texto, é o de permitir ter a informação necessária para caracterizar cada palavra ou frase. Facilita-se assim a detecção de padrões no texto que podem ser utilizados como conhecimento útil. Podem-se criar processos de extracção de conhecimento específico baseado em padrões do texto. Este tema da detecção e classificação de padrões, já foi estudada, p.ex. para o caso dos textos, onde se pretende detectar moradas [15]; e para um caso mais actual, onde se pretende determinar eventos num discurso falado [13]. Esta problemática está igualmente presente no estudo desta tese, pois os textos utilizados como amostras caracterizam-se pela abundância de expressões, que correspondem a padrões bem definidos. É salientado em [13], a importância que a correcta detecção desses padrões pode ter no processamento dos dados a posteriori.

Em [13] é feito um estudo onde se compara diversas técnicas de classificação de eventos. Estes eventos ocorrem no discurso falado, onde é necessário revelar a presença de elementos importantes que ocorrem no sinal sonoro do discurso. Num sistema de reconhecimento de voz, os eventos podem ser combinados para detectar telefones, palavras, frases, etc. A detecção destes eventos assenta, sobretudo num problema de classificação. Uma correcta abordagem a este problema é ter uma acção discriminativa, em que se diferenciam termos e padrões que ocorrem contiguamente. Para isto podem ser utilizados diversos algoritmos. Por exemplo [13] apresenta bons resultados na utilização de um modelo híbrido, baseado em HMM e redes neuronais. Estes algoritmos são utilizados como mecanismos de detecção, pois aprendem observando outros dados que definem eventos e padrões previamente tratados e conhecidos.

Apesar da utilização de dados textuais ser menos ambígua que a utilização de dados áudio (tal como os apresentados em [13]), existem pontos em comum com os dados textuais. Isto na medida em que, esta questão dos padrões/eventos que existem e que são importantes detectar e marcar, ocorre em ambos os cenários.

Seguindo ainda a linha dos padrões de palavras, importa referir a importância de conseguir classificar blocos lógicos. Algo que pode ser feito na sequência dessa classificação, é a criação de relações entre padrões. Em [23] é discutido o tema da criação de uma ontologia, baseada na relação entre as palavras de um dicionário. Para criar esta ontologia é necessário proceder à extracção das relações entre as palavras ou conceitos, criando estruturas organizadas que contêm essas relações. O objectivo final é gerar uma rede de relações semânticas. Para isso, inicialmente, são definidos alguns tipos de relações possíveis, entre palavras ou conceitos. Ao ser criada esta ontologia, com base nas relações detectadas, é possível utilizá-la para explorar e analisar o dicionário que serviu de base de conhecimento à ontologia.

Este estudo surge um pouco da ideia base do WordNet[6], em que as palavras, que constituem a base de conhecimento, estão ligadas a outras palavras que são sinónimos. Por exemplo, em [23] é proposto mais que uma ligação entre palavras que sejam sinónimos, são propostas outro tipo de ligações (*hiperonimo_de*(PalavraA, PalavraB), *causador_de*(PalA, PalB), *parte_de*(PalA, PalB), *meio_para*(PalA, PalB), *local_de*(PalA, PalB)). Este tipo de ligações permite obter uma rede de palavras ligadas por meio de diversas relações. Este tema foi abordado no âmbito desta tese. De facto, podem ser criadas regras (secção 4.2) que relacionam as palavras para auxiliar no processo de desambiguação das mesmas. A possibilidade de criar modelos, que nos permitam inferir sobre a base de conhecimento contendo as relações entre as palavras e os contextos (secção 3.7), é também extremamente relevante. Juntamente com os resultados apresentados em [22], e tal como será discutido na secção 4.2, os resultados experimentais apresentados nesta tese fundamentam a possibilidade de desenvolver ainda mais a base de conhecimento.

[6] Página Web: http://wordnet.princeton.edu/

A questão da relevância das palavras é um aspecto que pode ser importante, pois nos sistemas actuais normalmente existe sempre uma sobrecarga de informação. Assim, a capacidade de filtrar e discernir informações de maior relevância é importante. O estudo efectuado em [10] foca a análise e a recuperação da coesão referencial nos sumários, que utilizam a escolha de palavras (simples ou compostas) mais relevantes do texto para compor o sumário. O objectivo geral é enriquecer os sumários, analisando co-referências. Para isso é analisado o texto fonte, à procura de expressões que referenciem determinadas palavras contempladas no sumário. Desta forma é possível enriquecer os sumários gerados de forma exaustiva (gerado somente com base nas palavras mais relevantes) e manter uma coesão referencial dos sumários. Este tipo de possibilidades vão para além dos objectivos desta tese, no entanto pode-se relacionar com um modelo de *Data Warehouse* (secção 3.7). Para este caso, é importante ter uma forma de obter as palavras mais relevantes dos artigos, mantendo sempre a coesão referencial. Assim é possível obter um modelo multidimensional com fortes ligações entre os factos e as dimensões que constituem a base de conhecimento (secção 3.7).

## 2.3 Conceitos relevantes a nível da Aplicabilidade

### 2.3.1 Portal *econstroi*

Relativamente ao Portal, na Figura 2.1 pode-se ver o fluxo dos processos que funcionam no portal. Na criação do processo é definida a lista de artigos desejada pela entidade Compradora (1); segue-se a selecção de fornecedores, para os quais a lista de artigos (articulado, como exemplificado na Figura 2.3) será enviada (2); todos os fornecedores seleccionados recebem uma notificação da existência de uma nova consulta (3); essa consulta é analisada pelos fornecedores (4), que dão preços aos artigos e aos serviços requisitados (5); estas novas listas de artigos cotadas pelos Fornecedores são analisadas pelos Compradores (6), de modo a estes escolherem a oferta que consideram

melhor; finalmente, depois de escolhida a melhor oferta, é comunicado o Fornecedor seleccionado (7, 8).

Este é o caso mais simples que pode ocorrer, visto que é possível os Fornecedores proporem listas de artigos variantes, permitindo os Compradores escolherem mais do que uma proposta para adjudicar.

Tendo em conta este cenário, e considerando o tema desta tese, verifica-se que os serviços em que se podem aplicar as técnicas de *Text Mining* são principalmente o (1) e o (5), da Figura 2.1. Estes serviços são os que envolvem a escrita dos artigos, das suas descrições e das suas cotações. É nestes módulos que se estuda a eventual incorporação das técnicas propostas nesta tese. Este estudo é facilitado pela disponibilidade dos módulos de análise e teste existentes na arquitectura SOA do portal *econstroi*. Estes módulos são utilizados de modo obter algumas indicações sobre a utilidade das técnicas aqui propostas neste contexto de negócio.


Figura 2.1 – Funcionamento do mercado electrónico *econstroi*.

A Figura 2.2 mostra o actual método de criação de um novo artigo. O campo "Designação" corresponde à descrição do artigo. É desta forma que se pode adicionar artigos ao articulado, cujo exemplo é mostrado na Figura 2.3.

Importa referir que para adicionar um artigo é necessário que tenha sido criado pelo menos um capítulo no articulado, visto que os artigos são associados a um capítulo (Figura 2.3). Posteriormente pode ser associada uma categoria aos artigos, caso o utilizador pretenda.

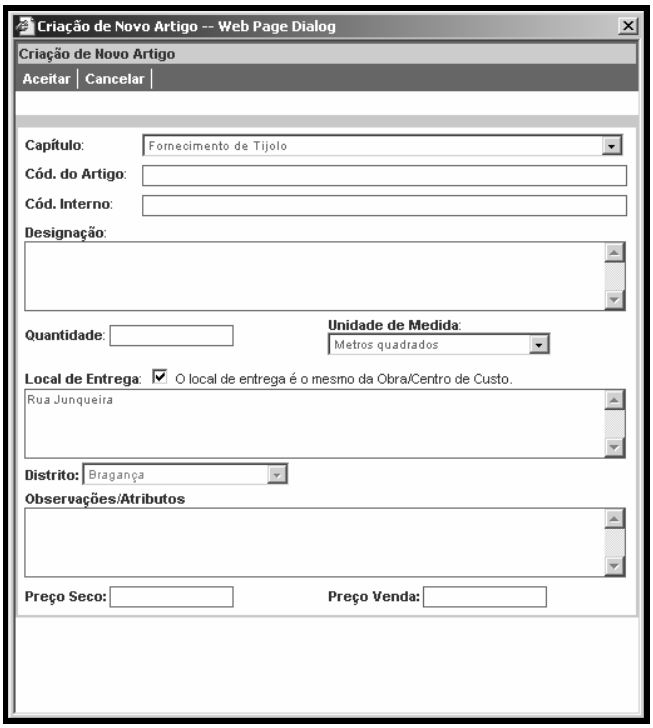



Figura 2.2 – Janela de criação de novo artigo.

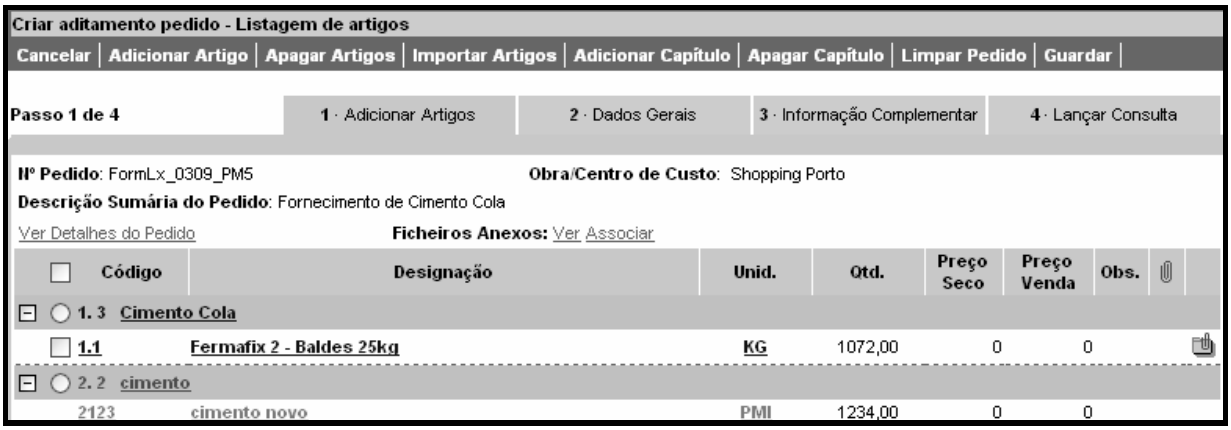



Figura 2.3 – Listagem dos artigos (articulado).

### 2.3.2 Aplicabilidade

A aplicabilidade que é referida no contexto desta tese está, por uma lado, relacionada com capacidade de expor determinadas funcionalidades ao utilizador final, e por outro, relacionada com a capacidade de desenvolver aplicações modulares que consigam interagir entre si.

Esta ideia é importante, visto que são abordados temas e tecnologias diferentes, cujo papel individual de cada um tem a sua importância, e somente com a contribuição de todas estas aplicações podemos chegar um resultado final válido. No caso desta tese, integra-se ferramentas de anotação morfossintáctica desenvolvidas em Prolog, com textos que foram extraídos da base de dados e tratados a nível textual, para poderem ser analisados, sendo ainda necessário extrair informação desses textos anotados de uma forma coerente e ajustada às necessidades do utilizador, tendo em conta o contexto dos textos analisados relativo a temas relacionados com a construção civil e materiais de construção.

### 2.3.3 Conceito B2B

Aliada a esta noção de aplicabilidade está o conceito B2B, que tem uma forte componente focada na extensibilidade, visto que a sua principal vantagem é a possibilidade de interacção a nível transaccional e aplicacional por parte de empresas e entidades.

A grande evolução tecnológica que tem havido a nível empresarial faz com que haja actualmente um elevado fluxo de troca de informação entre as empresas. Um grande impulsionador para o sucesso do conceito B2B são os *Web Services*, que permitem um elevado nível de abstracção, relativamente à troca de informação que é necessária realizar entre as diversas empresas que interagem entre si.

Devido à crescente exigência a nível da quantidade de serviços que as empresas usam e ao crescimento das relações empresariais existe a necessidade de por vezes, adaptar os *Web Services* a algo mais específico para o problema empresarial (*Business Services Networks*), e se que possa tornar numa plataforma capaz de suportar uma grande rede global de serviços [31]. Estes serviços são disponibilizados pelas empresas, podendo ser privados ou públicos, e consumidos pelas outras empresas. Desta forma é criado um modelo de negócio centrado numa rede aberta, onde as empresas podem ser construídas com base nos serviços disponibilizados por outras, criando assim um processo muito orientado ao crescimento individual de cada empresa, que implicitamente resulta no crescimento global do conjunto de empresas.

Esta rede de empresas propicia uma descentralização das ligações entre elas, obrigando a que haja uma integração entre todas as aplicações que fazem parte do ambiente criado [19].

Tendo em mente estes conceitos, que têm a capacidade de alterar a forma de comunicar e os processos a nível empresarial, surge a necessidade de existirem ontologias que providenciem uma terminologia que possa ser usada e entendida por todos os elementos [7]. Para isso é necessário haver repositórios de informação, havendo uma camada de abstracção, baseada em *Web Services*, que permita um fácil acesso, para consulta e enriquecimento, do conhecimento organizado segundo a ontologia. É comum existir mais do que uma ontologia partilhada, sendo ainda mais necessária a preocupação

em interligar todos estes repositórios de conhecimento, de modo a disponibilizar a informação como um todo.

### 2.3.4 Arquitectura Orientada aos Serviços

Comum a todos estes conceitos referidos, encontra-se a Arquitectura Orientada aos Serviços (SOA, *Service Oriented Architecture*) que define determinadas regras a seguir, de modo a que, no caso específico em estudo, os componentes da base de dados, a aplicação de análise morfológica, a aplicação de anotação morfossintáctica, as DCGs e a base de conhecimento consigam interagir correctamente e exista uma grande capacidade de disponibilização dessas funcionalidades, individuais ou como um todo, para que se possa acoplar a outros módulos aplicacionais que já existem (ver secção 2.3.1 deste capítulo) ou que possam surgir. A arquitectura SOA define igualmente conjuntos de métricas que devem ser aplicados aos diversos serviços, para que seja possível analisar e testar da sua utilidade e performance. Pretende-se aproveitar estes indicadores para validar a utilidade de integrar o estudo aqui proposto no portal *econstroi*.

Esta arquitectura é um conceito relativamente recente, que é impulsionado pela necessidade da melhoria no relacionamento existente entre as diversas áreas que suportam tecnologicamente os negócios e os negócios propriamente ditos. Assim, surge da necessidade e da conveniência do enfoque nos serviços, como objectivo final. Este conceito veio reforçar a necessidade de fortalecer o enfoque no cliente final e tornar a gestão de serviços como uma actividade produtiva, que gerasse valor à empresa [4]. A orientação ao serviço apresenta-se como a solução final para a organização ser ágil, sendo portanto, necessário que outras áreas da organização se adaptem (pessoas, processos e tecnologia) à metodologia SOA, de forma a poder haver uma agilidade global [8].

### 2.3.5 Gramática e análise sintáctica e semântica

As gramáticas de cláusulas definidas (DCGs) [24], são utilizadas para extrair uma representação lógica relativamente a determinados padrões de etiquetas que ocorrem nos textos. Neste sentido foram desenvolvidas várias regras gramaticais que possibilitam a extracção de conhecimento com padrões pré-determinados, e relevantes no contexto do portal *econstroi*. Na secção 3.3, apresentam-se já alguns primeiros exemplos.

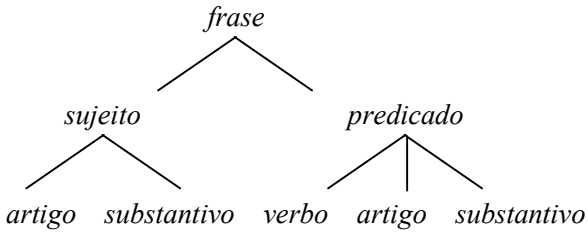
Para melhor se perceber o funcionamento das DCGs, é necessário compreender que uma gramática é uma especificação formal da estrutura das frases permitidas numa linguagem.

A forma mais simples de definir uma gramática é especificar um conjunto de símbolos *terminais*, que denotam palavras da linguagem, um conjunto de símbolos *não terminais*, que denotam os componentes das frases, e um conjunto de regras de produção, que expandem símbolos *não terminais* numa sequência de símbolos *terminais* e *não terminais*. É importante referir que a gramática deve ter um símbolo *não terminal* inicial, ou seja, o lado esquerdo de uma regra de produção tem sempre um único símbolo *não terminal*, enquanto o lado direito pode ter símbolos *terminais* e *não terminais* [24].

De seguida exemplifica-se esquematicamente a estrutura de uma gramática que define um fragmento da língua portuguesa. Optou-se por criar um exemplo fictício, sem qualquer relação com o tema da construção civil, de modo a que seja o mais simples possível. Nesta secção o intuito é demonstrar a construção mais básica possível de uma DCG. Um exemplo, com estrutura similar ao apresentado nesta secção, mas cujo tema relaciona-se com a construção civil pode ser consultado na secção 3.3 do capítulo 3.

No exemplo seguinte os símbolos *terminais* são **o**, **gato**, **rato** e **caçou**, sendo os restantes símbolos *não terminais*. A regra de produção *frase* → *sujeito predicado* estabelece que uma frase é composta por um sujeito seguido de um predicado. A regra *substantivo* → **gato** | **rato** estabelece que um substantivo pode ser a palavra "gato" ou "rato".

*frase* → *sujeito predicado*
*sujeito* → *artigo substantivo*
*predicado* → *verbo artigo substantivo*
*artigo* → **o**
*substantivo* → **gato** | **rato**
*verbo* → **caçou**




Árvore sintáctica do exemplo da esquerda

Em alternativa, esta árvore sintáctica poderia ser representada da seguinte forma:

*Frase* → *Sintagma_Nominal Sintagma_Verbal*
*Sintagma_Nominal* → *Determinante Nome*
*Sintagma_Verbal* → *Verbo Sintagma_Nominal*
*Determinante* → **o**
*Nome* → **gato** | **rato**
*Verbo* → **caçou**

Com a definição desta gramática pode-se reconhecer e validar a frase "o gato caçou o rato", seguindo duas abordagens: sintetizar a frase a ser reconhecida aplicando as regras de produção de forma progressiva ou derivando o símbolo inicial da gramática, a partir da frase a ser reconhecida, aplicando as regras de produção de forma regressiva. Um exemplo destes dois cenários, respectivamente, é demonstrado de seguida.

*frase*
→ *sujeito predicado*
→ *artigo substantivo predicado*
→ **o** *substantivo predicado*
→ **o gato** *predicado*
→ **o gato** *verbo artigo substantivo*
→ **o gato caçou** *artigo substantivo*
→ **o gato caçou o** *substantivo*
→ **o gato caçou o rato**

**o gato caçou o rato**
→ *artigo* **gato caçou o rato**
→ *artigo substantivo* **caçou o rato**
→ *sujeito* **caçou o rato**
→ *sujeito verbo* **o rato**
→ *sujeito verbo artigo* **rato**
→ *sujeito verbo artigo substantivo*
→ *sujeito predicado*
→ *frase*

O Prolog é uma linguagem bastante eficiente e intuitiva para o processamento de linguagem natural [24], e tem embutido uma implementação das DCGs. Assim podemos escrever gramáticas que podem ser utilizadas tanto para o reconhecimento, como para a geração de frases, de forma automática. A implementação da gramática definida anteriormente utilizando a sintaxe Prolog, e a notação DCG, é exemplificada de seguida.

```
frase --> sujeito, predicado.
sujeito --> artigo, substantivo.
predicado --> verbo, artigo, substantivo.
artigo --> [15].
substantivo --> [gato] | [rato].
verbo --> [caçou].
```

Tendo a gramática definida pode-se validar automaticamente qualquer frase, para verificar se está de acordo com a gramática definida, da seguinte forma:

```
?- frase([o, gato, caçou, o, rato],[]).
Yes


?- frase([o, gato, caçou, rato],[]).
No
```

No caso de se pretender gerar automaticamente frases, pode-se fazer "perguntas" à gramática e obter os resultado por *backtracking*, como se mostra de seguida.

```
?- frase(Frase, []).
Frase = [o, gato, caçou, o, gato] ;
Frase = [o, gato, caçou, o, rato] ;
Frase = [o, rato, caçou, o, gato] ;
Frase = [o, rato, caçou, o, rato]
No
```

Outra possibilidade é a obtenção da árvore sintáctica que gera a gramática, sendo necessário implementar a DCG da seguinte forma:

```
frase(frase(S, P)) --> sujeito(S), pred(P).
sujeito(sujeito(Art,Sub)) --> artigo(Art), subst(Sub).
pred(pred(Ver,Art,Sub)) --> verbo(Ver),artigo(Art),subst(Sub).
artigo(artigo(o)) --> [15].
subst(subst(gato)) --> [gato].
subst(subst(rato)) --> [rato].
verbo(verbo(caçou)) --> [caçou].
```

Assim, pode-se validar uma frase, obtendo a caracterização dos elementos da frase:

```
?- frase(Tree, [o,gato,caçou,o,rato],[]).
Tree   =   frase(sujeito(artigo(o),   subst(gato)),   pred(verbo(caçou),
artigo(o), subst(rato))) ;
No
```

No caso de se pretender obter as combinações possíveis para frases válidas, tendo em conta a gramática:

```
?- frase(Tree, Frase, []).
Tree   =   frase(sujeito(artigo(o),   subst(gato)),   pred(verbo(caçou),
artigo(o), subst(gato))),
Frase = [o, gato, caçou, o, gato] ;
Tree   =   frase(sujeito(artigo(o),   subst(gato)),   pred(verbo(caçou),
artigo(o), subst(rato))),
Frase = [o, gato, caçou, o, rato] ;
Tree   =   frase(sujeito(artigo(o),   subst(rato)),   pred(verbo(caçou),
artigo(o), subst(gato))),
Frase = [o, rato, caçou, o, gato] ;
Tree   =   frase(sujeito(artigo(o),   subst(rato)),   pred(verbo(caçou),
artigo(o), subst(rato))),
Frase = [o, rato, caçou, o, rato]
No
```

Isto suporta a opção tomada em utilizar as DCGs para extrair conhecimento de textos marcados, visto que desta forma pode-se facilmente definir gramáticas específicas para o problema em estudo e validar ou gerar frases a partir de bases de conhecimento estruturadas.

No capítulo 3 são mostrados alguns exemplos, que ilustram a utilização das DCGs na extracção de conhecimento de textos anotados com determinadas etiquetas.

# 3. Estrutura dos dados e extracção de conhecimento

Neste capítulo são explicados em detalhe os dados textuais, usados para demonstrar a aplicação das técnicas de processamento de língua natural estudadas nesta tese. Estas técnicas correspondem essencialmente à geração de uma base de conhecimento bem estruturada, à anotação das palavras do texto com etiquetas genéricas e específicas e à geração e reconhecimento de frases com base em gramáticas e regras de produção.

## 3.1 Estrutura dos textos

Como foi explicado na secção 2.3.1, existe uma interface própria para a criação de um articulado, onde é inserida uma descrição por cada artigo que é adicionado a esse articulado. Essa descrição é o que caracteriza o artigo pretendido, que poderá ser algo mais que um simples produto. Pode-se descrever as propriedades de um determinado produto ou os serviços pretendidos que estejam inerentes a esse produto. Esta informação textual e descritiva é inserida numa base de dados relacional. Assim, existe uma tabela que contém a listagem de todos os artigos, estando estes indexados a diferentes pedidos por parte das entidades compradoras.

Foi copiada uma amostra destes dados, sem haver qualquer relação com alguma entidade ou utilizador, de forma a obter um conjunto de textos. Para tal, foi exportado da base de dados um ficheiro XML com a listagem de 2275 descrições de artigos.

## 3.2 Estrutura dos dados e formato TXT/2

Inicialmente os dados textuais da amostra de descrição de artigos não tinham nenhuma formatação específica, sendo tratados como simples frases.

Para que se pudesse analisar os textos quanto à sua forma sintáctica foi necessário dividir cada frase, que corresponde a uma descrição de um artigo, obtendo assim as palavras separadas. Este processo de divisão das palavras foi efectuado utilizando uma aplicação, referida na secção 3.6, que permite a importação e exportação de dados em ficheiros XML. Depois de carregados os dados, esta aplicação permite manipular esses dados, fazendo a *estruturação dos dados* textuais carregados, oferecendo funcionalidades de *segmentação de frases* e *segmentação de palavras*, atribuindo automaticamente chaves identificadoras únicas (id's) às frases e às palavras separadas. Esta separação de palavras é feita com base nos caracteres de separação de palavras (espaço, vírgula, ponto, etc.), conforme descrito na secção 2.2. Esta segmentação pode ser auxiliada por Expressões Regulares, que podem definir uma determinada palavra complexa (exemplo: "1,20 m", "Ref. #03", etc.), podendo assim agrupar palavras e considerá-las como uma só.

***Estrutura inicial dos dados***

*Normalização dos dados:* Os dados originais, relativos aos textos a serem estudados para posterior anotação sintáctica, estão armazenados numa base de dados *SQL Server* de uma forma pouco estruturada. Com base nisto, é efectuado o carregamento dos mesmos de forma a estruturar os dados, ficando com as frases, que descrevem os artigos. De seguida (Tabela 3.1) é ilustrado, como exemplo, um objecto resultante da normalização que é efectuada aos dados não estruturados.

| sentenceId | sentenceStr |
|---|---|
| 1 | Fornecimento de mosaico hidráulico de 0,30x0,30 m. |
| 2 | Fornecimento de mosaico cerâmico de 0,40x0,40 m polido. |
| 3 | Soleira S1 com 0,36 m de largura. |
| ... | ... |

Tabela 3.1 – Normalização dos dados em frases.

***Estrutura de cada frase***

*Segmentação de palavras:* Depois dos dados estarem estruturados em frases coerentes, é necessário efectuar a segmentação, ficando as palavras separadas. Parte deste processo [30] foi analisado na secção 2.2, de trabalho relacionado. No caso específico deste trabalho foram utilizadas Expressões Regulares para definir o que é considerado como uma palavra, visto que o sistema automaticamente separa as palavras pelos símbolos mais básicos: . | , | ? | ! | “ ”.

Na Tabela 3.2 ilustra-se este tipo de segmentação (para a frase 1 e 3 da Tabela 3.1), tendo em conta que foi adicionada a seguinte Expressão Regular:

```
1. ((\d)+((\.|,)(\d)+)?(x)(\d)+((\.|,)(\d)+)?( )?(mm|cm|dm|m))
```

Esta Expressão Regular caracteriza a medida de uma área, sendo definida por um valor numérico (decimal ou não), seguido e um “x”, seguido novamente por um valor numérico (decimal ou não), terminando com as unidades (*m*, *dm*, *cm* ou *mm*).

Importa referir que a medida “0,36 m” não é definida por esta Expressão Regular, acabando por ser separada segundo os caracteres básicos de separação. O campo **`regularExpId`** da tabela indica qual a Expressão Regular que define a palavra, sendo 0 o valor indicativo de que a palavra não se encontra definida pela Expressão Regular.

| sentenceID | sentenceWordId | sentenceWord | regularExpId |
|---|---|---|---|
| 1 | 1 | Fornecimento | 0 |
| 1 | 2 | de | 0 |
| 1 | 3 | mosaico | 0 |
| 1 | 4 | hidráulico | 0 |
| 1 | 5 | de | 0 |
| 1 | 6 | 0,30x0,30 m | 1 |
| 1 | 7 | . | 0 |
| 3 | 1 | Soleira | 0 |
| 3 | 2 | S1 | 0 |
| 3 | 3 | com | 0 |
| 3 | 4 | 0 | 0 |
| 3 | 5 | , | 0 |
| 3 | 6 | 36 | 0 |
| 3 | 7 | m | 0 |
| 3 | 8 | de | 0 |
| 3 | 9 | largura | 0 |
| 3 | 10 | . | 0 |

Tabela 3.2 – Segmentação de frases.

***Formato TXT/2***

De forma a ter as frases com as palavras separadas num ficheiro de texto, foi criado o formato TXT/2 [20], com capacidade para conter etiquetas genéricas e palavras compostas. Desta forma é igualmente possível efectuar a análise sintáctica directamente aos textos que estejam neste formato.

Este é um formato lógico, implementado com base no SWI-Prolog, o que faz com que exista uma maior facilidade na utilização de linguagens de representação lógica para manipulação destes textos. Nomeadamente, as DCGs do Prolog, que são utilizadas neste trabalho, como uma ferramenta para expressão de gramáticas e sua aplicação.

Segundo este formato, as frases estão divididas em palavras, havendo estruturas de palavras aglomeradas numa lista Prolog. Esta lista representa todo o conhecimento necessário para o processamento do texto.

A Figura 3.3 mostra a estrutura de um TXT/2, que representa uma frase ou um texto, contendo várias palavras ($wd_i$ … $wd_n$). Cada uma das palavras é caracterizada por vários atributos, salientando-se os atributos que indicam a palavra (wd) e a etiqueta dessa palavra (tag – para etiquetas morfossintácticas). Cada uma destas palavras wd pode ter associada uma palavra composta cw, que representa uma composição dessa mesma palavra. Esta cw (palavra composta) é uma lista de wd's (palavras), que podem ter igualmente outras cw's (palavras compostas) associadas.

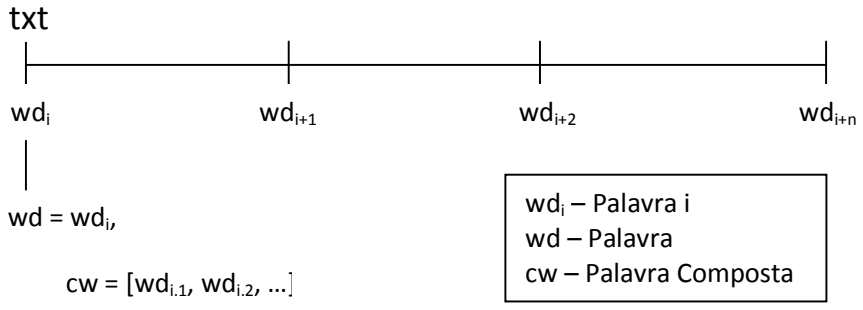



Figura 3.3 – Estrutura do TXT/2.

Nas secções 3.4 e 3.5 são mostrados e explicados exemplos de frases neste formato.

### 3.3 Anotação dos dados pelo NeSy Tagger com etiquetas genéricas

Os textos, depois de correctamente separados em palavras e codificados no formato TXT/2, seguiram para um processo de anotação morfossintáctica, utilizando o NeSy Tagger [14]. Nesta anotação foi utilizada uma rede neuronal previamente treinada com base nos textos do CETEMPublico [28].

Desta forma obteve-se uma classificação para todas as palavras, com etiquetas genéricas. Destas etiquetas destacam-se as seguintes: N (Nome), ADJ (Adjectivo), PROP (Nome próprio).

A Figura 3.5 mostra um texto, que corresponde a uma descrição de um artigo, e onde se pode ver a anotação feita pelo NeSy Tagger. Para cada palavra (wd) temos a classificação dada directamente pelo léxico (lex) e a anotação automática (atag), que tem em conta possíveis ambiguidades do anterior.

### 3.4 Anotação dos dados com etiquetas específicas

Para se tirar maior partido das DCGs foram criadas algumas etiquetas específicas (Tabela 3.4) para melhor extrair conhecimento especializado dos textos marcados. Na Figura 3.5 são mostrados alguns textos, meramente ilustrativos, anotados com as seguintes etiquetas específicas definidas:

| Etiqueta Específica | Descrição da Etiqueta | Exemplo de palavra |
|---|---|---|
| PROD | Produto | ex. “mosaico” |
| CA_PROD | Caracteriza o Produto | ex. “hidráulico” |
| CO_PROD | Constitui o Produto | ex. “lâminas” |
| SERV | Serviço | ex. “assentamento” |
| MED_PROD | Medida do Produto | ex. “0,30x0,30” |
| UNMED_PROD | Unidade de Medida do Produto | ex. “cm” |

Tabela 3.4 – Listagem das etiquetas específicas.

```
txt([16], [
     w([1, wd='Fornecimento', regex=0, tag='SERV']),
     w([2, wd='e', regex=0, atag=[['KC', 0.999359]], tag='PREP']),
     w([3, wd='assentamento', regex=0, tag='SERV']),
     w([4, wd='de', regex=0, atag=[['PREP', 0.999955]], tag='PREP']),
     w([5, wd='mosaico', regex=0, tag='PROD']),
     w([6, wd='hidráulico', regex=0, tag='CA_PROD']),
     w([7, wd='com', regex=0, atag=[['PREP', 0.99839]], tag='PREP']),
     w([8, wd='0,30x0,30', regex=1, tag='MED_PROD']),
     w([9, wd='m', regex=0, tag='UNMED_PROD']),
     w([10,wd='.', regex=0, atag=[['PTO', 0.999785]]]),[]]).
```

Figura 3.5 (a) – Texto anotado com etiquetas *(tags)*.

```
txt([5], [
     w([1, wd='Lâminas', regex=0, tag='CO_PROD']),
     w([2, wd='de', regex=0, atag=[['PREP', 0.999955]]]),
     w([3, wd='mármore', regex=0, tag='PROD']),
     w([4, wd='polido', regex=0, tag='CA_PROD']),
     w([5, wd='com', regex=0, atag=[['PREP', 0.99839]]]),
     w([6, wd='3 cm', regex=8, tag='MED_PROD']),
     w([7, wd='de', regex=0, atag=[['PREP', 0.999955]]]),
     w([8, wd='espessura', regex=0, atag=[['N', 0.997685]]),
     w([9, wd='.',regex=0, atag=[['PTO', 0.999785]]]),[]]).
```

Figura 3.5 (b) – Texto anotado com etiquetas *(tags)*.

***Criação das etiquetas específicas***

Estas etiquetas foram criadas com o intuito de poder classificar os textos de uma forma mais semântica, de modo a poder diferenciar as palavras numa vertente mais específica ao tema da construção.

No total da amostra foram identificadas manualmente um conjunto de regras minimamente representativas das frases em estudo. Assim, tendo em conta alguma análise que foi efectuada a todas as frases da amostra total de textos da construção, foram identificadas determinadas palavras-chave mais utilizadas nesses textos. Desta forma foi

obtida uma lista de palavras mais importantes e específicas da área da construção, onde se verificou que muitas dessas palavras correspondiam a produtos da construção, características dos produtos, constituintes dos produtos, pedidos de serviços e mão-de-obra e medidas (comprimentos, áreas e volumes). Assim, foram criadas seis etiquetas específicas que conseguissem caracterizar estes temas com algum nível de detalhe.

***Anotação com base nas etiquetas específicas***

Depois de ter o levantamento das palavras-chave e de ter a especificação das etiquetas específicas a serem utilizadas, procedeu-se à anotação das frases com as etiquetas específicas. Para tal foi criado um dicionário específico com estas palavras, que foi anotado manualmente com as etiquetas específicas. Para criar este dicionário foi feito um levantamento das palavras mais usadas em todas as frases da amostra, removendo as *stopwords*. Alguns casos em que as palavras ocorriam em frases com categorias diferentes, foi atribuída mais que uma etiqueta específica na entrada dessa palavra no dicionário.

Importa notar que algumas palavras deste dicionário têm duas etiquetas específicas associadas, criando assim uma possível fonte de ambiguidade. Isto acontece visto que foram detectadas palavras que podiam ter mais do que um sentido, consoante o contexto da frase. Este tema é estudado e explicado em detalhe na secção 4.1.

Assim, a cada palavra do texto que está presente no dicionário específico, é inserida a etiqueta específica correspondente ao campo "*tag*" da palavra.

Como se pode ver na Figura 3.6 cada linha contém uma palavra e a sua etiqueta específica correspondente. De notar que no caso de existir mais do que uma etiqueta específica correspondente, considera-se numa primeira fase a etiqueta mais provável, que é a que está à cabeça da lista. No entanto, conforme será discutido no capítulo 4, esta anotação poderá ser sujeita a revisão.

```
dic('assentamento', [['SERV', 121]], 121).
dic('montagem', [['SERV', 153]], 153).
dic('moleanos', [['CA_PROD', 235]], 235).
dic('mosaico', [['PROD', 153]], 153).
dic('plástico', [['PROD', 21], ['CA_PROD', 5]], 26).
(...)
```

Figura 3.6 – Excerto do dicionário sobre artigos de construção (notação SWI-Prolog).

Outro aspecto que importa referir na anotação, para além deste dicionário específico, é a indicação das etiquetas específicas por parte das Expressões Regulares (campo "*regex*" no formato TXT/2). Cada uma destas expressões regulares tem associada uma etiqueta específica, permitindo assim a anotação de palavras complexas que não se encontram no dicionário específico, mas que foram detectadas pelas Expressões Regulares definidas pelo utilizador.

Estas etiquetas têm um papel muito importante na extracção de conhecimento mais específico dos textos marcados, como é demonstrado na secção 3.5.

## 3.5 Gramáticas e regras de produção

***Exemplo de uma gramática específica simples para extrair conhecimento***

Tendo em conta as regras de produção da gramática da Figura 3.7, aplicada aos textos marcados da Figura 3.5, podemos gerar frases (Figura 3.8) com base nos Produtos (PROD), nas Medidas dos Produtos (MED_PROD) e nas Unidades de Medida dos Produtos (UNMED_PROD).

A gramática que se segue é meramente ilustrativa das funcionalidades que podem ser aplicadas aos textos anotados. Esta gramática extrai composições de palavras para formar uma frase válida pela gramática, consultando para isso os textos anotados que estão disponíveis e obedecendo à formatação exemplificada na Figura 3.5.

```
artigoConst --> nome_art, prep, dimensao, medida.
nome_art -->[X], {txt(_,Ls), member(w(List), Ls), member((wd=X), List),
            member((tag='PROD'), List)}.
prep -->    [X], {txt(_,Ls), member(w(List), Ls), member((wd=X), List),
            member((tag='PREP'), List)}.
dimensao -->[X], {txt(_,Ls), member(w(List), Ls), member((wd=X), List),
            member((tag='MED_PROD'), List)}.
medida -->  [X], {txt(_,Ls), member(w(List), Ls), member((wd=X), List),
            member((tag='UNMED_PROD'), List)}.
```

Figura 3.7 – Gramática para extrair conhecimento (notação SWI-Prolog).

Na Figura 3.8 podemos ver que é possível validar se a frase pertence à gramática apresentada na Figura 3.7. Pode-se assim utilizar esta técnica para aplicação de testes ao sistema que se pretende integrar na aplicação SOA *econstroi*, como foi referido na secção 2.3.1.

```
?- artigoConst(['Mosaico', com, '0,30x0,30', cm],[]).
Yes
?- artigoConst(['Mosaico', com, '0,30x0,30', metros],[]).
No
```

Figura 3.8 – Validação de frase por parte da gramática.

A Figura 3.9 mostra uma chamada à gramática, onde se passa uma variável X para obtermos todos os resultados possíveis por *backtracking*. Neste caso, ao ser chamada a gramática da Figura 3.7, esta vai gerar as combinações que são válidas, de acordo com a gramática da Figura 3.7, tendo como base o texto anotado da Figura 3.5.

Desta forma, podemos também efectuar testes e analisar a aplicabilidade do sistema, tendo em conta a integração numa aplicação SOA, como foi referido no exemplo 3.8.

```
?- artigoConst(X,[]).
X = ['Mosaico', com, '0,30x0,30', m] ;
X = ['Mosaico', com, '0,30x0,30', cm] ;
X = ['Mosaico', com, '3', m] ;
X = ['Mosaico', com, '3', cm] ;
X = ['Mosaico', com, '0,30x0,30', m] ;
X = ['Mosaico', com, '0,30x0,30', cm] ;
X = ['Mosaico', com, '3', m] ;
X = ['Mosaico', com, '3', cm] ;
X = ['Mosaico', de, '0,30x0,30', m] ;
X = ['Mosaico', de, '0,30x0,30', cm] ;
X = ['Mosaico', de, '3', m] ;
X = ['Mosaico', de, '3', cm] ;
X = ['Pedra', com, '0,30x0,30', m] ;
X = ['Pedra', com, '0,30x0,30', cm] ;
X = ['Pedra', com, '3', m] ;
X = ['Pedra', com, '3', cm] ;
X = ['Pedra', com, '0,30x0,30', m] ;
X = ['Pedra', com, '0,30x0,30', cm] ;
X = ['Pedra', com, '3', m] ;
X = ['Pedra', com, '3', cm] ;
X = ['Pedra', de, '0,30x0,30', m] ;
X = ['Pedra', de, '0,30x0,30', cm] ;
X = ['Pedra', de, '3', m] ;
X = ['Pedra', de, '3', cm] ; No
```

Figura 3.9 – Geração de conhecimento com base numa gramática.

No caso de se pretender obter sugestões por parte da gramática impondo condicionantes, terá de se chamar o método da seguinte forma:

```
?- artigoConst([('Pedra')|S],[]).

S = [com, '0,30x0,30', m] ;
S = [com, '0,30x0,30', cm] ;
S = [com, '3', m] ;
S = [com, '3', cm] ;
S = [com, '0,30x0,30', m] ;
S = [com, '0,30x0,30', cm] ;
S = [com, '3', m] ;
S = [com, '3', cm] ;
S = [de, '0,30x0,30', m] ;
S = [de, '0,30x0,30', cm] ;
S = [de, '3', m] ;
S = [de, '3', cm] ; No
```

Figura 3.10 – Geração de conhecimento com base numa gramática, impondo condições.

***Regras de produção da gramática específica***

A estrutura das regras de produção da gramática que utiliza as etiquetas específicas é a seguinte:

| | |
|---|---|
| Regra 1 | PROD MED_PROD |
| Regra 2 | CO_PROD de PROD do tipo CA_PROD |
| Regra 3 | SERV de PROD do tipo CA_PROD com dimensões MED_PROD |
| Regra 4 | PROD do tipo CA_PROD com espessura de MED_PROD e acabamento CA_PROD |
| Regra 5 | SERV de PROD CA_PROD |
| Regra 6 | PROD CA_PROD com MED_PROD de espessura SERV CA_PROD |
| Regra 7 | SERV de CO_PROD de PROD |
| Regra 8 | SERV e SERV de PROD |

Tabela 3.11 – Regras de produção da gramática específica.

Tendo em conta as regras da Tabela 3.11, de seguida são mostrados exemplos de frases válidas para cada uma das regras de produção:

Regra 1. Portas 2.5x2.5 m.

Regra 2. Blocos de mármore do tipo polido.

Regra 3. Fornecimento de tijolo do tipo cerâmico com dimensões 0,30x0,30 m.

Regra 4. Varão do tipo vidraço com espessura de 3 mm e acabamento polido.

Regra 5. Aplicação de mosaico plástico.

Regra 6. Pedra moleanos com 4 cm de espessura acabamento amaciado.

Regra 7. Montagem de blocos de vedações.

Regra 8. Pintura e montagem de portas.

### *Exemplo de gramática complexa e árvore sintáctica*

Considerando agora a frase da Figura 3.12, e em que as etiquetas específicas criadas já se tornam insuficientes para classificar, de uma forma rica e coesa, a frase na sua totalidade. Podemos observar de seguida a classificação feita pelo NeSy Tagger (Figura 3.13), pelo LX-Suite7 (Figura 3.14) e a árvore sintáctica gerada pelo VISL[8] (Figura 3.15).

```
Fornecimento e assentamento de pavimento em soalho de madeira de pinho,
com 22 mm de espessura e 12 cm de largura, incluindo o preenchimento com
granulado de cortiça, afagamento e envernizamento.
```

Figura 3.12 – Frase que descreve o fornecimento de um produto e a sua aplicação.

### *NeSy Tagger*

```
txt([16], [
  w([1, wd='fornecimento',lex=[['N', 0.910463]],atag(rnd)=[['N', 0.99482]], regex=0, tag='SERV']),
  w([2, wd='e', lex=[['KC', 0.999359]], atag(rnd)=[['ADJ', 0.12701]], regex=0]),
  w([3, wd='assentamento',lex=[['N', 0.910463]],atag(rnd)=[['N', 0.99504]], regex=0, tag='SERV']),
  w([4, wd='de', lex=[['PRP', 0.995771]], atag(rnd)=[['PRP', 0.9999]], regex=0, tag='PRP']),
  w([5, wd='pavimento', lex=[['N', 0.864942]], atag(rnd)=[['N', 0.99672]], regex=0, tag='PROD']),
  w([6, wd='em', lex=[['PRP', 0.99839]], atag(rnd)=[['PRP', 0.99969]], regex=0]),
  w([7, wd='soalho', lex=[['N', 0.864942]], atag(rnd)=[['N', 0.99677]], regex=0]),
  w([8, wd='de', lex=[['PRP', 0.995771]], atag(rnd)=[['PRP', 0.99968]], regex=0, tag='PRP']),
  w([9, wd='madeira', lex=[['N', 0.864942]], atag(rnd)=[['N', 0.9967]], regex=0, tag='PROD']),
  w([10, wd='de', lex=[['PRP', 0.995771]], atag(rnd)=[['PRP', 0.99967]], regex=0, tag='PRP']),
  w([11, wd='pinho', lex=[['N', 0.910463]], atag(rnd)=[['N', 0.99731]], regex=0]),
  w([12, wd=',', lex=[['PU', 0.999825]], atag(rnd)=[['PU', 0.98234]], regex=0]),
  w([13, wd='com', lex=[['PRP', 0.998487]], atag(rnd)=[['PRP', 0.99919]], regex=0, tag='PRP']),
  w([14, wd='22 mm', cw=[
     w([14.1, wd='22', lex=[['PROP', 0.641617]], regex=8, atag(rnd)=[['PROP', 0.69134]]]),
     w([14.2, wd=mm, lex=[['PROP', 0.641617]], regex=8, atag(rnd)=[['PROP', 0.93256]]]),[]],
     lex=[['PROP', 0.641617]], regex=8, tag='MED_PROD']),
  w([15, wd='de', lex=[['PRP', 0.995771]], atag(rnd)=[['PRP', 0.99981]], regex=0, tag='PRP']),
  w([16, wd='espessura',lex=[['N',0.910463]],atag(rnd)=[['N', 0.99557]], regex=0, tag='CA_PROD']),
  w([17, wd='e', lex=[['KC', 0.999359]], atag(rnd)=[['V-PCP', 0.07653]], regex=0]),
  w([18, wd='12 cm', cw=[
     w([18.1, wd='12', lex=[['PROP', 0.641617]], regex=8, atag(rnd)=[['PROP', 0.69134]]]),
     w([18.2, wd=cm,lex=[['PROP',0.641617]],regex=8,atag(rnd)=[['PROP',0.93256]]]),[]],
     lex=[['PROP', 0.641617]], regex=8, tag='MED_PROD']),
  w([19, wd='de', lex=[['PRP', 0.995771]], atag(rnd)=[['PRP', 0.99995]], regex=0, tag='PRP']),
```

[7] http://lx-suite.net

[8] http://beta.visl.sdu.dk/visl/pt/parsing/automatic/trees.php (Projecto Floresta Sintá(c)tica)

```
 w([20, wd='largura',lex=[['N', 0.910463]], atag(rnd)=[['N', 0.99082]], regex=0, tag='CA_PROD']),
 w([21, wd=',', lex=[['PU', 0.999825]], atag(rnd)=[['PU', 0.99126]], regex=0]),
 w([22, wd='incluindo', lex=[['V', 0.995261]], atag(rnd)=[['EC', 0.18261]], regex=0]),
 w([23, wd='o', lex=[['DET', 0.962651]], atag(rnd)=[['EC', 0.05677]], regex=0]),
 w([24, wd='preenchimento', lex=[['N', 0.910463]], atag(rnd)=[['N', 0.95998]], regex=0]),
 w([25, wd='com', lex=[['PRP', 0.998487]], atag(rnd)=[['PRP', 0.99975]], regex=0, tag='PRP']),
 w([26, wd='granulado', lex=[['V', 0.727763]], atag(rnd)=[['N', 0.09712]], regex=0]),
 w([27, wd='de', lex=[['PRP', 0.995771]], atag(rnd)=[['PRP', 0.99963]], regex=0, tag='PRP']),
 w([28, wd='cortiça', lex=[['N', 0.910463]], atag(rnd)=[['N', 0.99421]], regex=0]),
 w([29, wd=',', lex=[['PU', 0.999825]], atag(rnd)=[['PU', 0.99126]], regex=0]),
 w([30, wd='afagamento', lex=[['N', 0.910463]], atag(rnd)=[['N', 0.99251]], regex=0]),
 w([31, wd='e', lex=[['KC', 0.999359]], atag(rnd)=[['IN', 0.07075]], regex=0]),
 w([32, wd='envernizamento',lex=[['N',0.910463]],atag(rnd)=[['N',0.98282]],regex=0, tag='SERV']),
 w([33, wd='.', lex=[['PU', 0.99962]], atag(rnd)=[['PU', 0.98931]], regex=0]),
[]]).
```

Figura 3.13 – Frase do exemplo 3.12 analisada pelo NeSy Tagger e no formato TXT/2.

### *Análise LX-Suite*

```
<p><s>
 Fornecimento/FORNECIMENTO/CN#ms e/CJ assentamento/ASSENTAMENTO/CN#ms de/PREP
 pavimento/PAVIMENTO/CN#ms em/PREP soalho/SOALHO/CN#ms de/PREP madeira/MADEIRA/CN#gs de/PREP
 pinho/PINHO/CN#ms ,*//PNT com/PREP 22/DGT mm/?/V#? de/PREP espessura/ESPESSURA/CN#fs e/CJ 12/DGT
 cm/?/V#? de/PREP largura/LARGURA/CN#fs ,*//PNT incluindo/INCLUIR/V#ger o/DA#ms
 preenchimento/PREENCHIMENTO/CN#ms com/PREP granulado/GRANULAR,GRANULADO/PPA#ms de/PREP
 cortiça/CORTIÇA/CN#fs ,*//PNT afagamento/AFAGAMENTO/CN#ms e/CJ
 envernizamento/ENVERNIZAMENTO/CN#ms ./PNT
</s></p>
```

Figura 3.14 – Frase do exemplo 3.12 analisada pelo LX-Suite.

*Árvore sintáctica utilizando etiquetas genéricas*

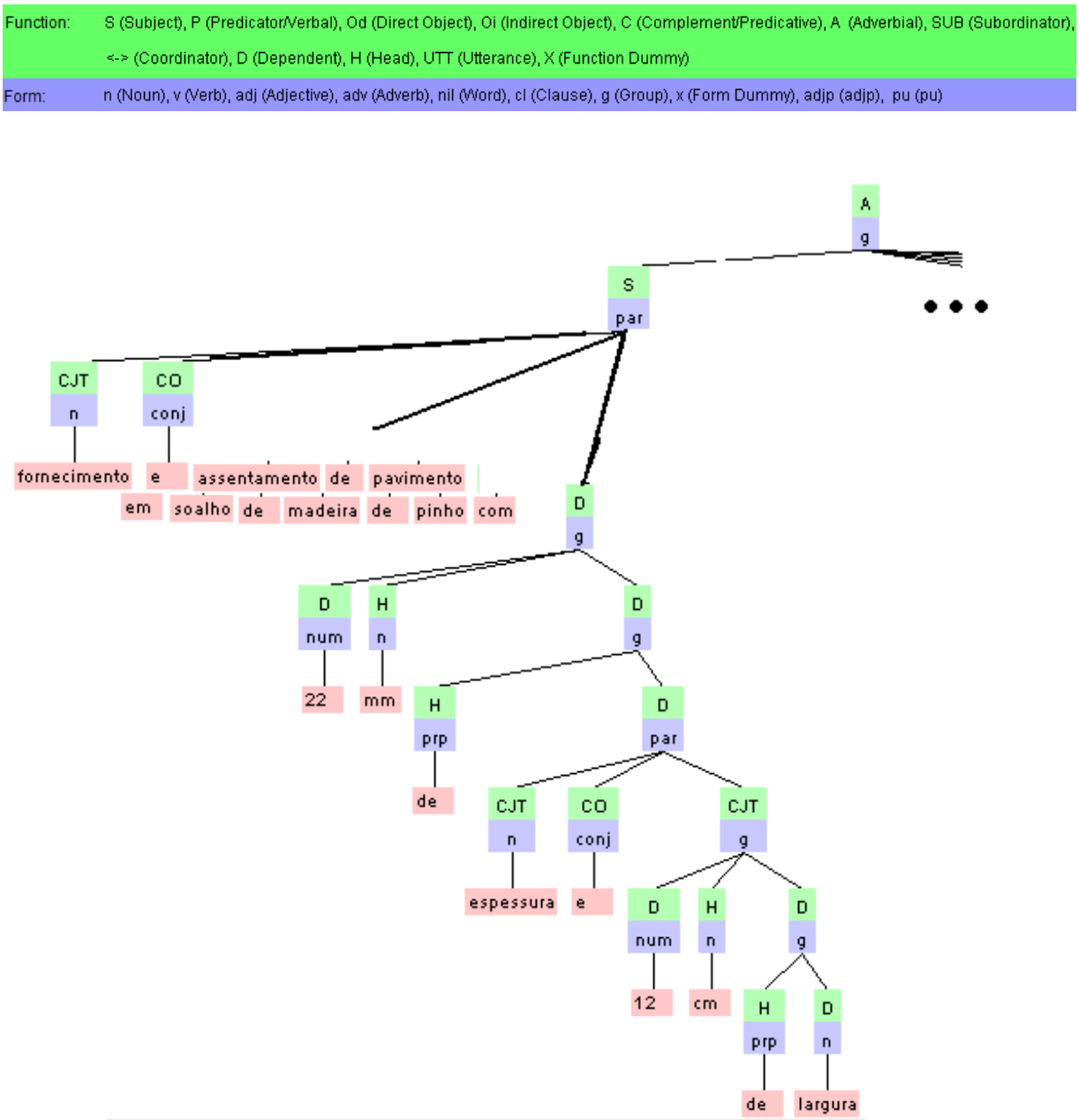



Figura 3.15(a) – Árvore sintáctica da frase do exemplo 3.8 criada pelo VISL.

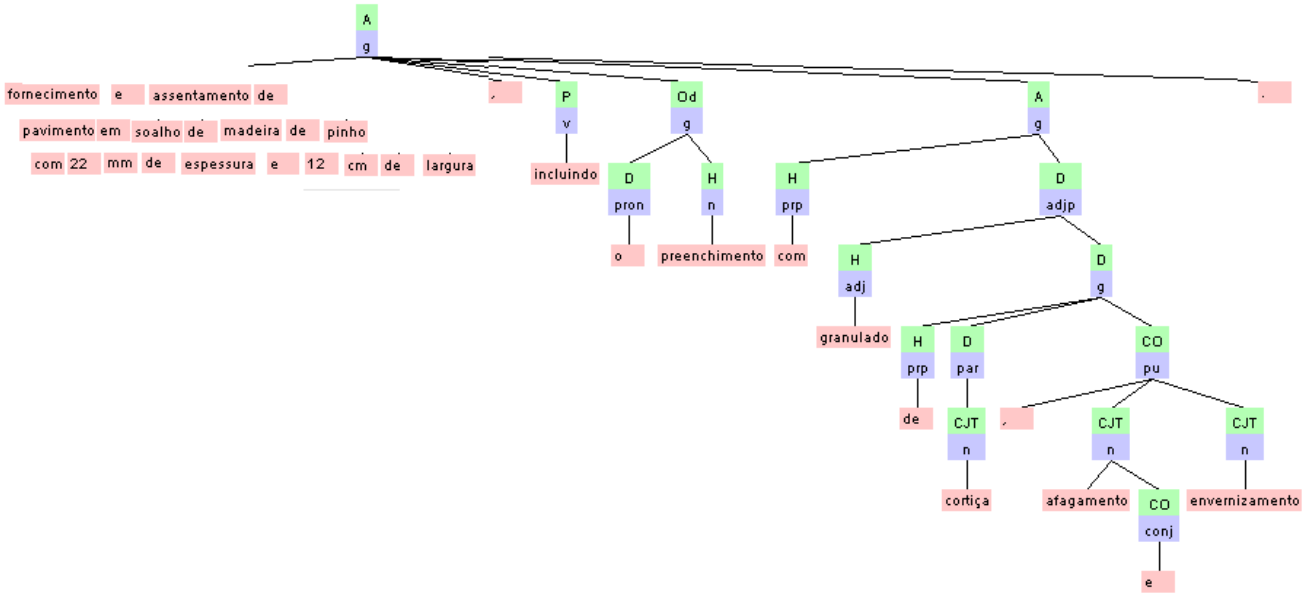



Figura 3.15(b) – Árvore sintáctica da frase do exemplo 3.8 criada pelo VISL.

Verifica-se que a classificação é bastante semelhante entre os três classificadores utilizados. De notar que o VISL dá-nos a possibilidade de analisar a anotação de uma forma gráfica, segundo uma árvore sintáctica.

Apesar da complexidade e especificidade da frase da Figura 3.8, de seguida é apresentada uma gramática, na forma de DCG, que representa esta frase. Para isso foram utilizando símbolos não terminais que caracterizam determinadas fracções da frase e símbolos terminais que servem para validar as palavras da frase. Com base nesta DCG foi gerada uma árvore sintáctica, que permite ter uma visão da configuração da frase face aos elementos da área da construção criados para o efeito, ou seja, estes elementos servem para reconhecer pequenos fragmentos da frase.

*Árvore sintáctica utilizando etiquetas específicas para a área da construção*

```
frase(frase(P, T)) --> infoProduto(P), infoTrabalhos(T).

infoProduto(infoProduto(T,PL1,PC,PL2,MP)) --> tarefa(T), pL(PL1),
produtoCompleto(PC), pL(PL2), medidaDoProduto(MP).

infoTrabalhos(infoTrabalhos(PL1,PL2,MO1,PL3,TC,MO2,PL4,MO3)) --> pL(PL1), pL(PL2),
maoDeObra(MO1), pL(PL3), tarefaCaracterizada(TC), maoDeObra(MO2), pL(PL4),
maoDeObra(MO3).

tarefaCaracterizada(tarefaCaracterizada(CP,PL,P)) --> constituiProduto(CP), pL(PL),
produto(P).

tarefa(tarefa(P)) --> provisao(P).
tarefa(tarefa(MO)) --> maoDeObra(MO).
tarefa(tarefa(P,PL,MO)) --> provisao(P), pL(PL), maoDeObra(MO).

produtoCompleto(produtoCompleto(P,PL,MP)) --> produto(P), pL(PL),
materialDoProduto(MP).

materialDoProduto(materialDoProduto(P,PL,PC)) --> produto(P), pL(PL),
produtoCaracterizado(PC).

produtoCaracterizado(produtoCaracterizado(P,PL,CP)) --> produto(P), pL(PL),
caractProduto(CP).

medidaDoProduto(medidaDoProduto(M)) --> medida(M).
medidaDoProduto(medidaDoProduto(M1,PL,M2)) --> medida(M1), pL(PL), medida(M2).
medida(medida(N,U,PL,T)) --> numero(N), unidade(U), pL(PL), tipoMedida(T).
provisao(provisao(fornecimento)) --> [fornecimento].
maoDeObra(maoDeObra(assentamento)) --> [assentamento].
maoDeObra(maoDeObra(preenchimento)) --> [preenchimento].
maoDeObra(maoDeObra(afagamento)) --> [afagamento].
maoDeObra(maoDeObra(envernizamento)) --> [envernizamento].
produto(produto(pavimento)) --> [pavimento].
produto(produto(soalho)) --> [soalho].
produto(produto(madeira)) --> [madeira].
produto(produto(cortiça)) --> [cortiça].
constituiProduto(constituiProduto(granulado)) --> [granulado].
caractProduto(caractProduto(pinho)) --> [pinho].
tipoMedida(tipoMedida(espessura)) --> [espessura].
tipoMedida(tipoMedida(largura)) --> [largura].
pL(pL(e)) --> [5].
pL(pL(de)) --> [de].
pL(pL(em)) --> [em].
pL(pL(com)) --> [com].
pL(pL(incluindo)) --> [incluindo].
pL(pL(o)) --> [15].
numero(numero(22)) --> [28].
numero(numero(12)) --> [29].
unidade(unidade(mm)) --> [mm].
unidade(unidade(cm)) --> [cm].
```

Figura 3.16 – DCG específica para a frase estudada.

***Árvore sintáctica da frase, gerada com a DCG anterior***

```
Tree = frase(
  infoProduto(
    tarefa(
      provisao(fornecimento), pL(e), maoDeObra(assentamento)
    ),
    pL(de),
    produtoCompleto(
      produto(pavimento), pL(em),
      materialDoProduto(
        produto(soalho), pL(de),
        produtoCaracterizado(
          produto(madeira),
          pL(de),
          caractProduto(pinho)
        )
      )
    ),
    pL(com),
    medidaDoProduto(
      medida(
        numero(22),
        unidade(mm),
        pL(de),
        tipoMedida(espessura)
      ),
      pL(e),
      medida(
        numero(12),
        unidade(cm),
        pL(de),
        tipoMedida(largura)
      )
    )
  ),
  infoTrabalhos(
    pL(incluindo),
    pL(o),
    maoDeObra(preenchimento),
    pL(com),
    tarefaCaracterizada(
      constituiProduto(granulado),
      pL(de),
      produto(cortiça)
    ),
    maoDeObra(afagamento),
    pL(e),
    maoDeObra(envernizamento)
  )
)
```

Figura 3.17 – Árvore sintáctica gerada pela DCG.

Todas estas abordagens face às gramáticas são válidas, mas dependendo da finalidade e do objectivo de cada trabalho, cada uma pode ter mais vantagens face às outras.

Neste caso, verifica-se a maior proximidade entre a frase anotada pelo NeSy Tagger e a DCG criada, na medida em que o TXT/2 consegue reunir tanto a informação genérica como a específica. Deste modo o texto é facilmente manipulável em qualquer sentido, tanto numa vertente mais genérica como numa vertente específica, com a possibilidade, inclusive, de serem geradas árvores sintácticas, para validação ou geração de conhecimento.

## 3.6 Protótipo para geração e validação de frases

Nesta secção é mostrado um exemplo de aplicação baseado nos conceitos estudados no capítulo 3. É com base neste protótipo que foram feitos os ensaios experimentais do capítulo 4.

Na Figura 3.18 pode-se ver que o protótipo permite gerar frases com base num tema escolhido. Cada tema corresponde a uma gramática criada, que pode ter diversas regras de produção. Desta forma é possível escolher um tema e á medida que se vai escrevendo vão aparecendo sugestões para completar a frase. Estas sugestões, como são geradas pelas regras de produção da gramática, são automaticamente válidas.

Dependendo do tema escolhido, obtém-se estruturas de frases e palavras diferentes, visto estarem relacionadas com a gramática escolhida. Importa referir a possibilidade de obter a geração de frases com base nas gramáticas criadas com etiquetas genéricas, onde os resultados são em maior quantidade, mas não têm tanta qualidade. Este aspecto é aprofundado no capítulo 4.

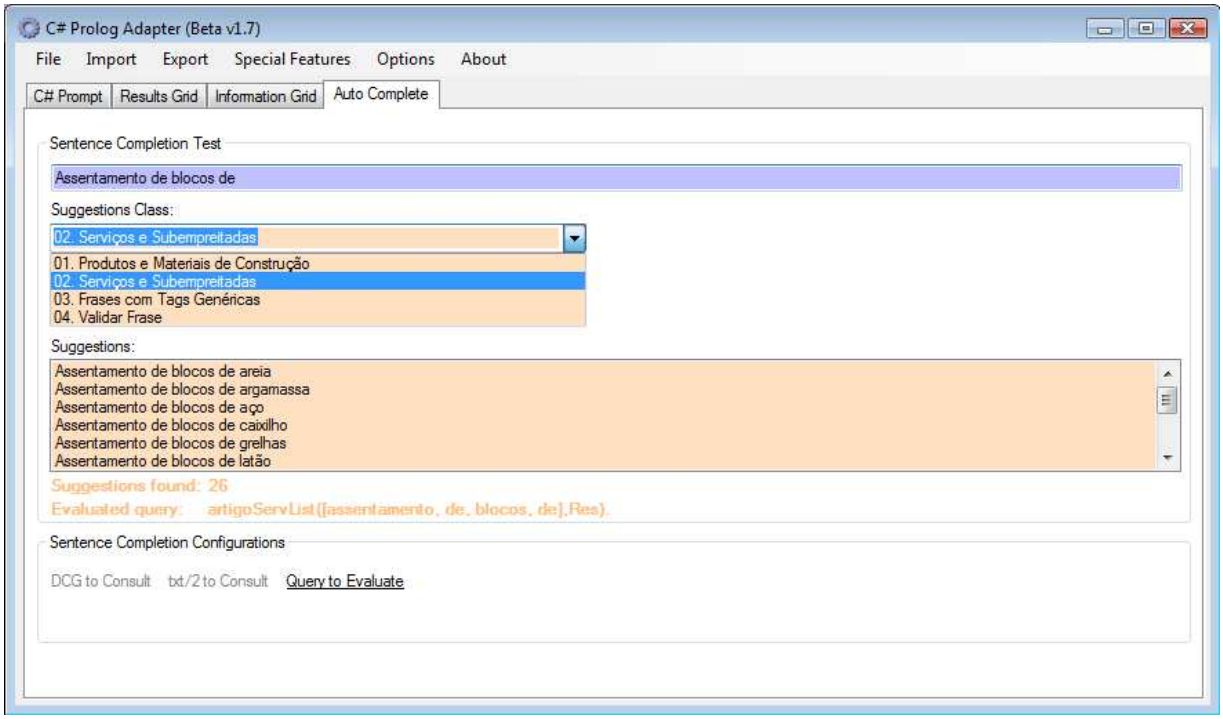



Figura 3.18 – Protótipo: Geração de frase.

Na Figura 3.19 pode-se ver que o protótipo permite reconhecer frases com base nas gramáticas definidas.

O processo baseia-se em fazer uma chamada a cada uma das regras gramaticais definidas. Ao ser feita esta chamada, passa-se a frase inserida pelo utilizador, na forma de uma lista Prolog com as palavras da frase separadas (ver na Figura 3.19 o campo *Evaluated query*). Desta forma as regras gramaticais vão tentar unificar a lista fornecida com as suas regras baseadas nas etiquetas. Assim que alguma das regras gramaticais sucede e não falha, é automaticamente retornada a informação do id da gramática que validou a frase. Sendo essa informação mostrada ao utilizador (ver na Figura 3.19 a zona *Suggestions*).

No caso da Figura 3.19, a primeira regra gramatical que validou a frase foi a **grm(7)**. Consultando a Tabela 3.11 da secção 3.5, podemos observar as regras que foram definidas para esta estrutura gramatical.

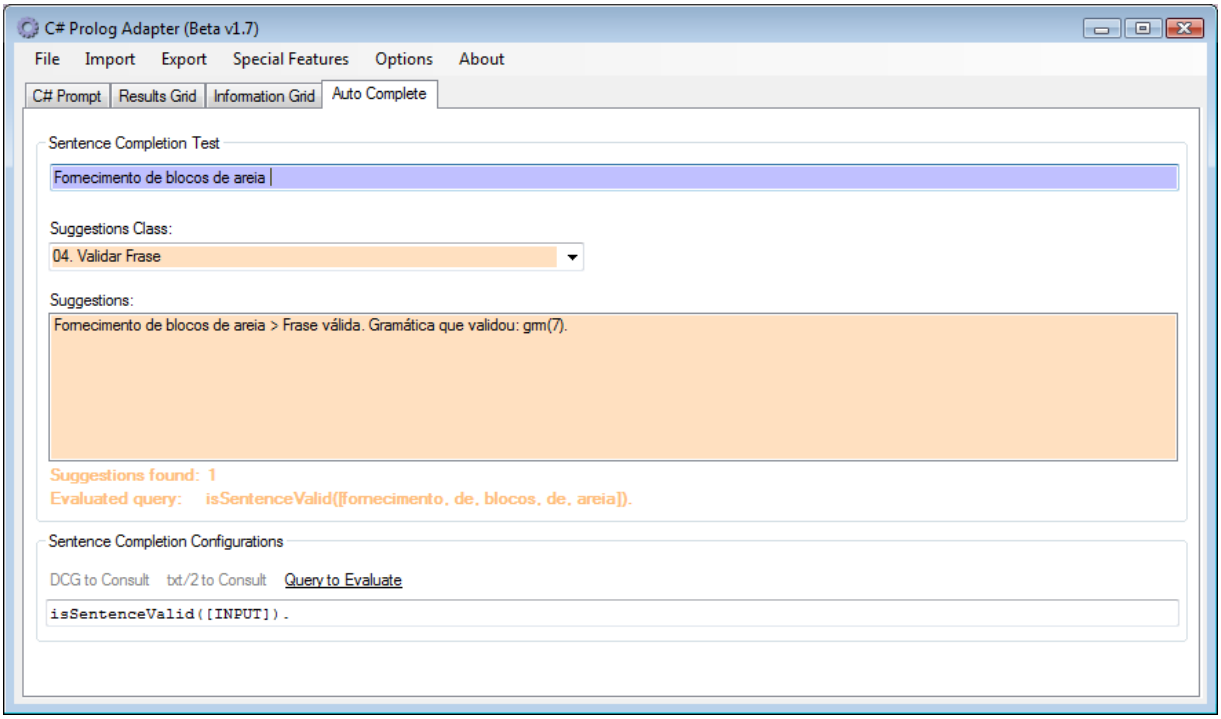



Figura 3.19 – Protótipo: Reconhecimento de frase.

## 3.7 Modelos para extracção de conhecimento

O processo de extracção de informação dos dados textuais, estudado e analisado nesta tese, permite obter dados bem estruturados, que podem posteriormente ser tratados de diversas formas. A Figura 3.20 mostra, de uma forma muito simplista, um conjunto de passos que permitem partir de um ponto onde os dados textuais estão completamente dispersos e não estruturados (1), passando por uma fase intermédia (2) onde os dados são analisados e estruturados segundo frases e palavras, até chegar a uma fase final onde a informação se pode tornar em conhecimento (3), conseguindo ser completamente absorvido esse conhecimento em prol de aplicações que nos transmitem informação útil e compreensível para o utilizador.

Na última fase (3) da Figura 3.20, é possível ver alguns tipos de aplicações que podem utilizar os dados estruturados obtidos na fase (2).

No caso (3a) é aproveitada a aplicação de gramáticas e regras de produção para gerar conhecimento mais estruturado e lógico com base nos dados textuais previamente tratados. Desta forma, são criadas regras que organizam construções de novos tipos de frases. Este aspecto é estudado e analisado na secção 3.6 e 4.3, da dissertação.

No caso (3b) é feito o uso de regras de validação, onde se valida a estrutura das frases submetidas e as próprias palavras. Isto com base em regras gramaticais e no conteúdo dos textos que foram tratados e estruturados. Um exemplo deste caso é aprofundado na secção 3.6 e 4.4, da dissertação.

O caso (3c) é discutido de uma forma mais detalhada nesta secção. Isto porque, para além dos casos descritos anteriormente, onde há uma preocupação em conseguir expor a informação de forma produtiva ao utilizador final (de modo a ser possível extrair conhecimento), é também importante, disponibilizar internamente toda a informação necessária sobre o funcionamento da organização. No fundo, é dar a possibilidade das empresas conhecerem melhor toda a informação que circula dentro delas, e assim utilizar melhor essa informação para ajudar nas tomadas de decisão. É nestes casos que se tem verificado a importância em adoptar arquitecturas de *Data Warehousing*. Estas arquitecturas são criadas como módulos externos, constituídos por modelos multidimensionais, que vão sendo alimentados com dados do sistema. De maneira geral,

as tabelas destes modelos representam transacções, acontecimentos, ou outras actividades que são utilizadas para medir o funcionamento e os resultados dos processos de negócio. A construção do *Data Warehouse* pode ainda ser complementada, com outra abordagem, relacionada com o processamento dos dados obtidos na fase (3), utilizando ferramentas de *Data Mining*. Este tipo de ferramentas permite a obtenção de conhecimento, a partir da informação existente.

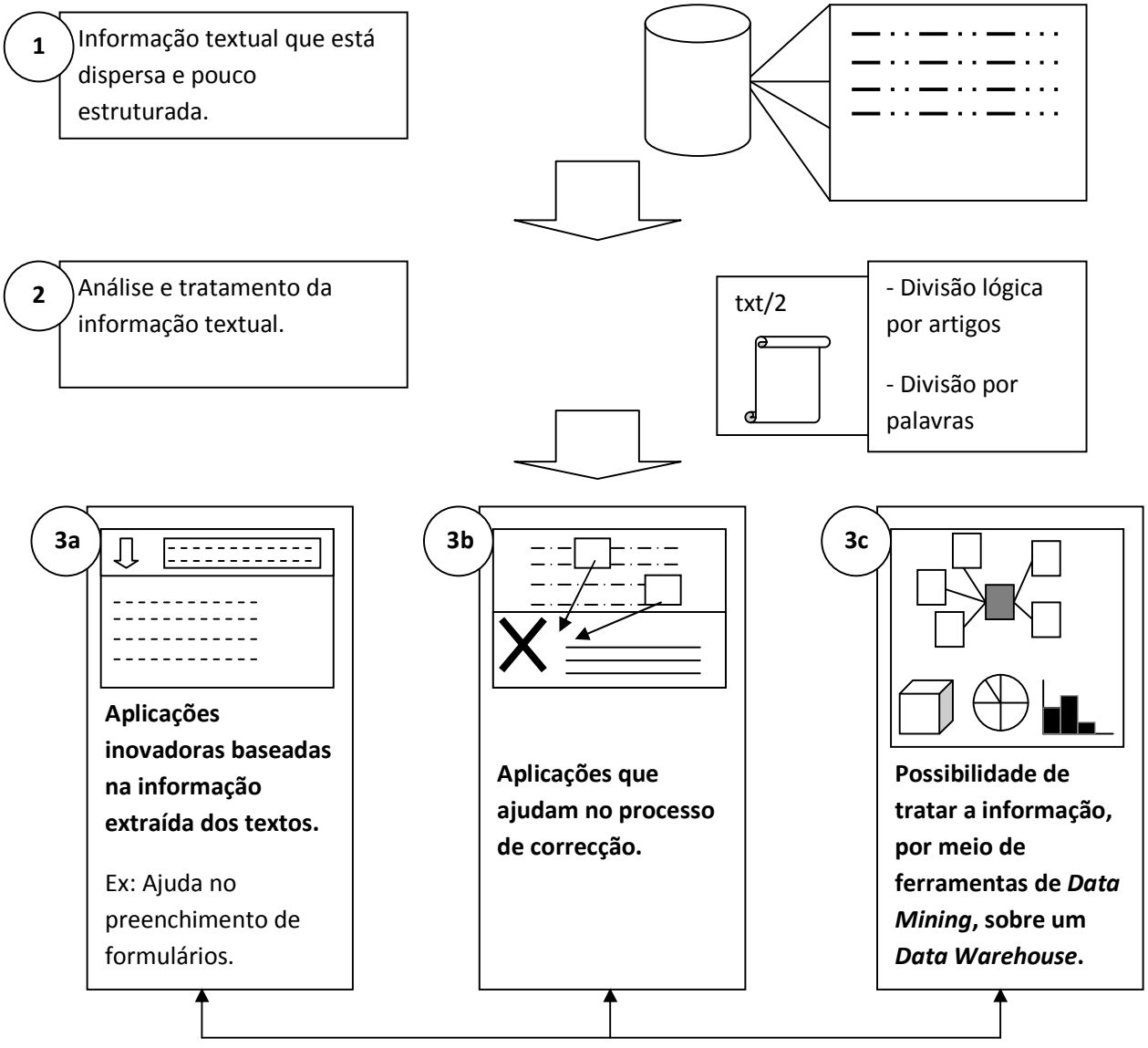



Figura 3.20 – Modelo geral do processo de extracção de informação e conhecimento.

### 3.7.1 *Data Warehouse*

Para o caso específico dos dados textuais utilizados neste estudo, e tendo em conta o modelo clássico de um *Data Warehouse*, é proposto aqui um modelo em estrela (Figura 3.22). Neste modelo existe uma tabela central (tabela de factos) que está ligada a diversas tabelas dimensionais (tabelas de dimensão).

Como se pode ver na Figura 3.23, a tabela central contém os artigos existentes. As tabelas de dimensão correspondem a informações obtidas através das etiquetas que marcam o texto e através dos dados presentes na tabela original de artigos.

Considerando o exemplo da Figura 3.21, onde é descrito um determinado artigo, para se perceber melhor a forma de obter os dados necessários para povoar as tabelas aqui descritas. Pode-se observar que o Artigo (tabela Artigo_fact) é constituído por um serviço de fornecimento (tabela Serviço_dim), um produto (tabela Produto_dim) mosaico do tipo pladur e medida 30 x 30 cm.

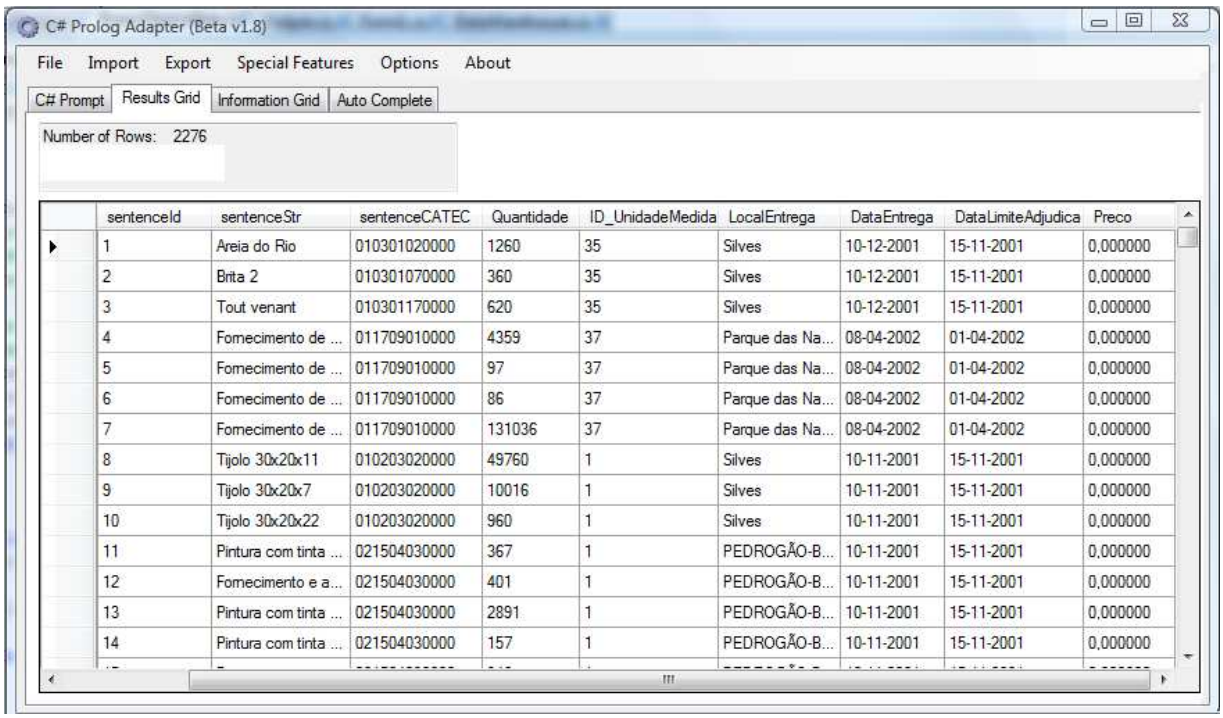
| Fornecimento | de | 100 caixas | de | mosaico | do tipo | pladur, | com | 30 x 30 cm. |
|---|---|---|---|---|---|---|---|---|
| SERV | | *QUANT*[9] | | PROD | | CA_PROD | | MED_PROD |

Figura 3.21 – Exemplo de descrição de um artigo.

As restantes tabelas (Categoria_dim, UnidadeMedida_dim, LocalEntrega_dim, DataEntrega-dim e DataLimiteAdjudicacao_dim) são preenchidas com base em dados retirados da tabela original de artigos, e não através da análise do texto da descrição desses artigos. A Figura 3.22 mostra a tabela original de artigos, onde se pode observar as descrições de cada artigo na coluna *sentenceStr* e os dados associados a esta descrição: *sentenceCATEC* (Categoria do artigo), *Quantidade* (Quantidade do artigo), *ID_UnidadeMedida* (Unidade de medida relativa à quantidade do artigo), *LocalEntrega* (Local de entrega do artigo), *DataEntrega* (Data de entrega do artigo),

[9] Etiqueta específica (QUANT: Quantidade) não contemplada no estudo desta tese. Está aqui descrita de forma meramente informativa.

*DataLimiteAdjudicacao* (Data limite para adjudicar o fornecimento desse artigo), *Preco* (Preço total do artigo).




| sentenceId | sentenceStr | sentenceCATEC | Quantidade | ID_UnidadeMedida | LocalEntrega | DataEntrega | DataLimiteAdjudica | Preco |
|---|---|---|---|---|---|---|---|---|
| 1 | Areia do Rio | 010301020000 | 1260 | 35 | Silves | 10-12-2001 | 15-11-2001 | 0,000000 |
| 2 | Brita 2 | 010301070000 | 360 | 35 | Silves | 10-12-2001 | 15-11-2001 | 0,000000 |
| 3 | Tout venant | 010301170000 | 620 | 35 | Silves | 10-12-2001 | 15-11-2001 | 0,000000 |
| 4 | Fornecimento de ... | 011709010000 | 4359 | 37 | Parque das Na... | 08-04-2002 | 01-04-2002 | 0,000000 |
| 5 | Fornecimento de ... | 011709010000 | 97 | 37 | Parque das Na... | 08-04-2002 | 01-04-2002 | 0,000000 |
| 6 | Fornecimento de ... | 011709010000 | 86 | 37 | Parque das Na... | 08-04-2002 | 01-04-2002 | 0,000000 |
| 7 | Fornecimento de ... | 011709010000 | 131036 | 37 | Parque das Na... | 08-04-2002 | 01-04-2002 | 0,000000 |
| 8 | Tijolo 30x20x11 | 010203020000 | 49760 | 1 | Silves | 10-11-2001 | 15-11-2001 | 0,000000 |
| 9 | Tijolo 30x20x7 | 010203020000 | 10016 | 1 | Silves | 10-11-2001 | 15-11-2001 | 0,000000 |
| 10 | Tijolo 30x20x22 | 010203020000 | 960 | 1 | Silves | 10-11-2001 | 15-11-2001 | 0,000000 |
| 11 | Pintura com tinta ... | 021504030000 | 367 | 1 | PEDROGÃO-B... | 10-11-2001 | 15-11-2001 | 0,000000 |
| 12 | Fornecimento e a... | 021504030000 | 401 | 1 | PEDROGÃO-B... | 10-11-2001 | 15-11-2001 | 0,000000 |
| 13 | Pintura com tinta ... | 021504030000 | 2891 | 1 | PEDROGÃO-B... | 10-11-2001 | 15-11-2001 | 0,000000 |
| 14 | Pintura com tinta ... | 021504030000 | 157 | 1 | PEDROGÃO-B... | 10-11-2001 | 15-11-2001 | 0,000000 |

Figura 3.22 – Tabela original com as informações completas sobre o artigo.

É com estes dados que se alimenta o modelo aqui proposto. O modelo proposto (Figura 3.23) é justificado tendo em conta os dados disponíveis. Por um lado, a partir da análise que é feita às descrições dos artigos (coluna sentenceStr, da Figura 3.22) e por outro aos dados associados. Deste modo, é possível caracterizar cada uma das tabelas apresentadas na Figura 3.23. As tabelas Produto_dim, Servico_dim e Artigo_fact são alimentadas somente com base na análise que é feita ao texto das descrições dos artigos. A tabela UnidadeMedida_dim é alimentada a partir dos dados originais da tabela de artigos, mas relaciona-se de uma forma directa com o texto relativo à descrição do artigo, pois determina a unidade das medidas que são descritas nesse texto. A tabela Categoria_dim contém as categorias de produtos. Esta tabela está directamente relacionada com o texto que constitui a descrição do artigo, permitindo, a nível de trabalho futuro, novos desenvolvimentos relativos à relação desta categoria com o tipo de

produtos descritos na descrição do artigo. As tabelas LocalEntrega_dim, DataEntrega_dim e DataLimiteAdjudicacao_dim são alimentadas com base nos dados da tabela de artigos e servem essencialmente para *Business Intelligence*. Isto porque, estas tabelas caracterizam temporalmente e espacialmente as descrições dos artigos, ou seja, definem o local onde deverá ser entregue o artigo e quando ele deve ser entregue.

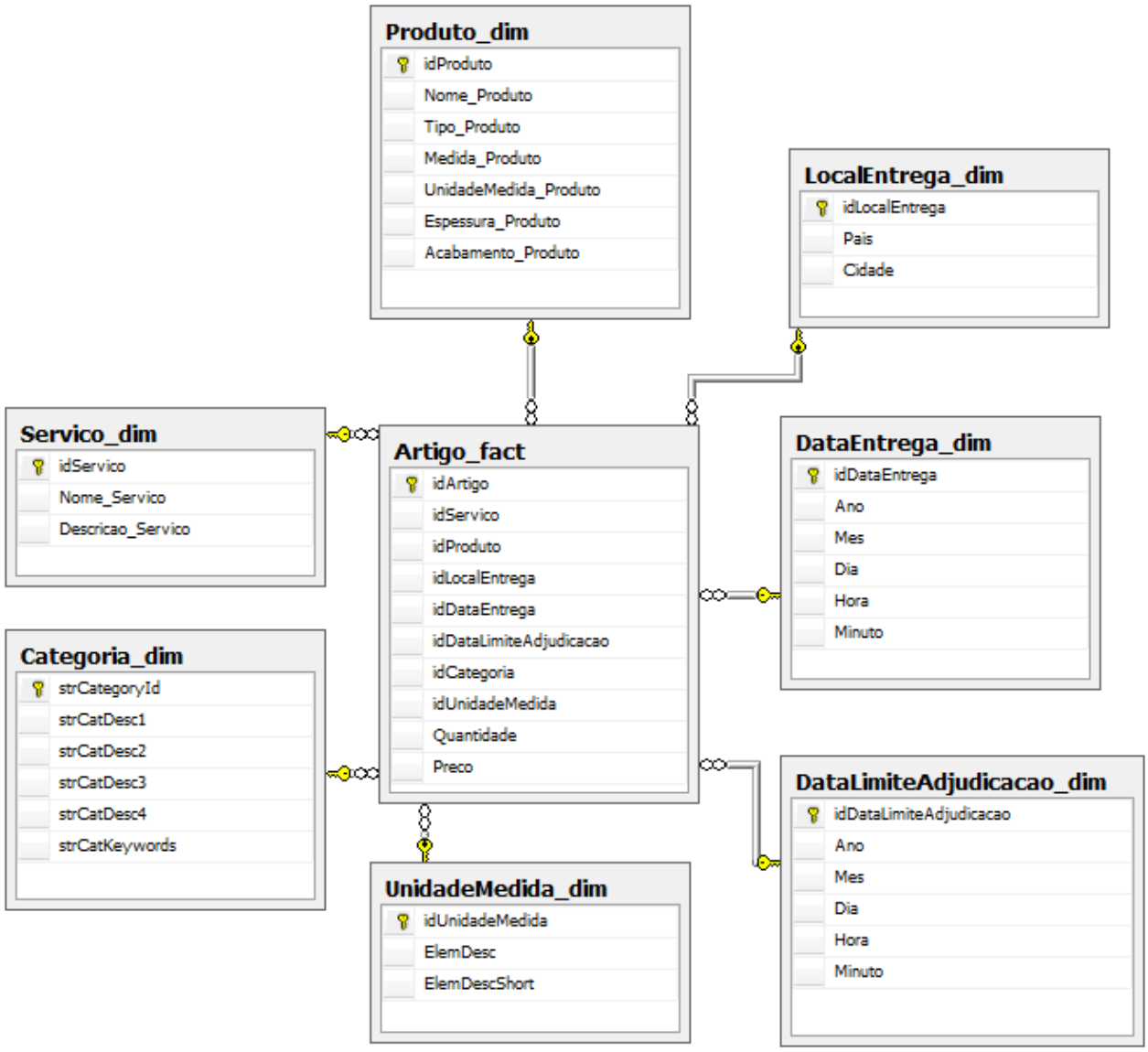



Figura 3.23 – Modelo relacional dos dados que constituem o artigo.

Na Tabela 3.24 é apresentado um exemplo da informação que pode estar contida na tabela Produto_dim. As descrições dos artigos que deram origem aos dados da Tabela 3.24 são as seguintes:

- **Pedra Moleanos** com **4 cm** de espessura, acabamento **amaciado**.
- **Pedra Moleanos** com **6 cm** de espessura, acabamento **bujardado** a pico fino.
- **Mosaico pitonado** com **30 x 30 x 1,4 cm**, cor amarelo.
- **Tijolo** de **Vidro 30 x 20 x 11 cm**.
- **Tijolo Vazado 0,4 x 0,3 x 0,05 m**.

A forma de detectar as várias informações relativas ao artigo, é ter em conta a anotação feita às palavras, com as etiquetas genéricas e específicas. Desta forma, é possível obter de forma directa algumas características da descrição do artigo. O nome do produto e a medida podem ser obtidos analisando as etiquetas PROD e MED_PROD , respectivamente. É necessário verificar igualmente se entre a ocorrência destas duas etiquetas, não aparecem outras que possam comprometer a correcta relação entre estas duas (p.ex. se ocorrer outro produto). Uma análise mais detalhada é apresentada no Capítulo 4.

**Produto_dim**

| idProduto (PK) | Nome_Produto | Tipo_Produto | Medida_Prod | Espessura_Prod | Acabamento_Prod |
|---|---|---|---|---|---|
| 1 | Pedra | Moleanos | | 4 cm | Amaciado |
| 2 | Pedra | Moleanos | | 6 cm | Bujardado |
| 3 | Mosaico | Pitonado | 30 x 30 x 1,4 cm | | |
| 4 | Tijolo | Vidro | 30 x 20 x 11 cm | | |
| 5 | Tijolo | Vazado | 0,4 x 0,3 x 0,05 m | | |

Tabela 3.24 – Tabela de dimensão Produto_dim com alguns dados de exemplo.

Com base nestas tabelas relacionais é possível saber, por ex., as medidas de todos os tijolos, que tipos de mosaicos existem, em que altura do ano se transacciona mais cimento ou mais madeira, etc. Note-se que esta informação não está disponível (directamente) na base de conhecimento da Vortal.

Com este tipo de modelo de dados consegue-se obter uma grande variedade de indicadores de desempenho do negócio, contribuindo para um aproveitamento do conhecimento contido nos dados textuais, que não existe actualmente.

### 3.7.2 *Implementação do Cubo*

De modo a tirar maior partido dos dados presentes nas tabelas descritas anteriormente, foi criado um Cubo, utilizando a ferramenta *Microsoft SQL Server Analysis Services* [10]. Na criação do Cubo foi definido que as tabelas Artigo_fact, DataEntrega_dim e DataLimiteAdjudicacao_dim fornecessem os atributos relativos às medidas (*Measures*, Figura 3.25); e que os parâmetros das tabelas Servico_dim, DataLimiteAdjudicacao_dim, DataEntrega_dim, LocalEntrega_dim, UnidadeMedida_dim, Produto_dim e Categoria_dim servissem de dimensões (*Dimensions*, Figura 3.25). A Figura 3.26 mostra as tabelas que concretizam esta descrição.

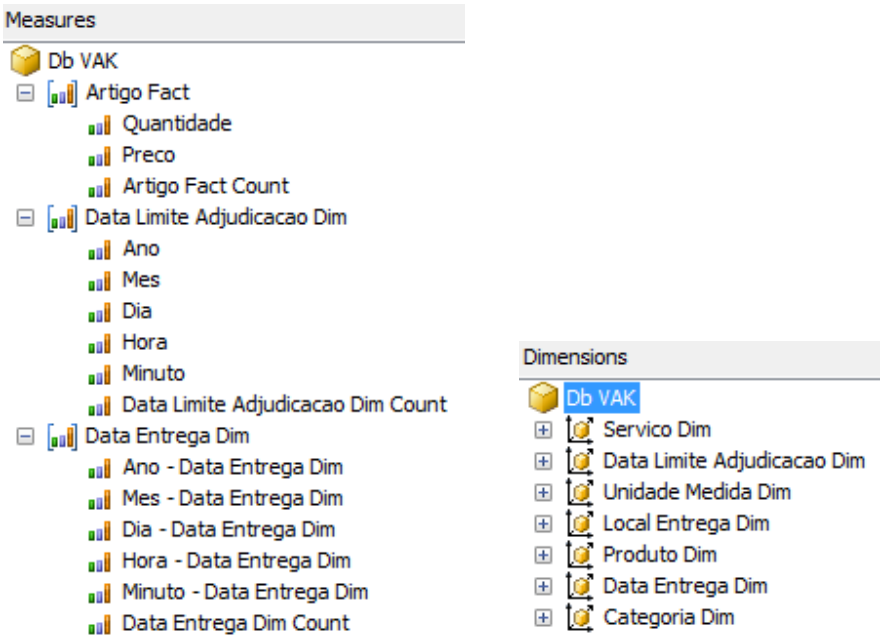



Figura 3.25 – Definição das Medidas e das Dimensões do Cubo.

[10] http://msdn.microsoft.com/en-us/library/ms175609(SQL.90).aspx

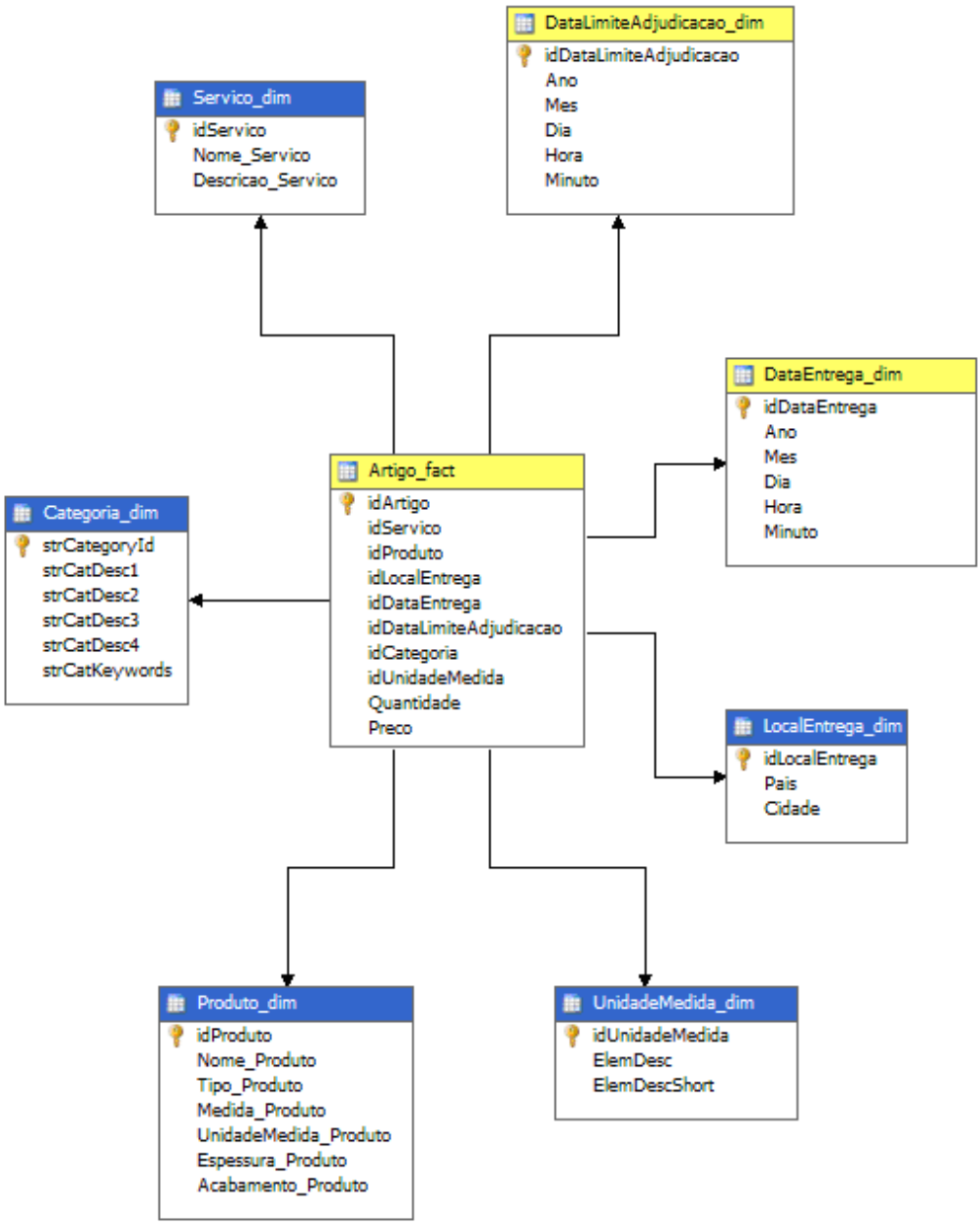



Figura 3.26 – Esquema multidimensional em forma de estrela, que constitui o Cubo.

Os dados utilizados para construir o Cubo são relativos à amostra de 2275 frases (descrições de artigos) que foi previamente analisada e classificada (ver Secção 3.2 e Capítulo 4 para melhor compreender a classificação que foi feita ao texto). Para cada descrição de um artigo foram analisadas as etiquetas (Figura 3.21), e com base nisso

preencheu-se as tabelas Servico_dim, Produto_dim e Artigo_dim. Para tal, foram considerados alguns factores, de modo a facilitar a extracção de informação a partir da descrição do artigo: numa descrição só se considera a existência de um Serviço, de um Produto, de uma característica desse Produto, de uma medida desse Produto e de um tipo de acabamento. Outra característica para que o artigo fosse contabilizado é ter um Produto classificado como tal. Isto significa que para cada descrição de um artigo só irá resultar um tuplo para a tabela Produto_dim e outro para a tabela Artigo_dim.

Depois do processamento das descrições de artigos obteve-se as seguintes estatísticas:

| | |
|---|---|
| Produto_dim | 534 linhas |
| Servico_dim | 25 linhas |
| Artigo_fact | 1418 linhas |

Nota: Podem existir na tabela Produto_dim produtos com nomes iguais, desde que algum dos outros parâmetros da tabela seja diferente.

Foram contabilizados 534 Produtos diferentes, tendo em conta as suas medidas, a sua espessura e o acabamento. Foram contabilizados 25 Serviços diferentes (p.ex. fornecimento, assentamento, aplicação, pintura, escavação, etc). Relativamente ao número de artigos, que efectivamente foram inseridos, o valor é de 1418 artigos. Isto significa que foram contabilizados 62% dos artigos totais da amostra. Este facto deve-se principalmente ao não reconhecimento de alguns produtos e à qualidade dos dados originais, que é discutido no Capítulo 4.

Tendo este modelo implementado, é possível fazer diversos tipos de questões no âmbito do *Business Intelligence*. É neste ponto que a informação que foi extraída dos textos, analisada e classificada, se pode tornar conhecimento útil para o utilizador.

Por exemplo, com o auxílio da tabela *Pivot*, é possível determinar a quantidade de tijolo transaccionado em cada um dos meses do ano de 2002, conforme mostra a Figura 3.27.

| | | Nome Produto | |
|---|---|---|---|
| | | tijolo | Grand Total |
| Ano | Mes | Artigo Fact Count | Artigo Fact Count |
| 2002 | 1 | 5 | 5 |
| | 2 | 4 | 4 |
| | 3 | 16 | 16 |
| | 6 | 23 | 23 |
| | 10 | 4 | 4 |
| | 11 | 1 | 1 |
| | Total | 53 | 53 |
| Grand Total | | 53 | 53 |

Figura 3.27 – Total de artigos, com tijolo, transaccionados em 2002.

Note-se, devido aos textos em estudo estarem restringidos apenas a um subconjunto de todas as transacções, existem muitos meses que não estão contemplados. Ainda assim, se assumíssemos a amostra como representativa, notaríamos que a grande maioria do tijolo transaccionado é relativo aos meses de Março e Junho.

A Figura 3.28 mostra como é possível determinar que categorias foram atribuídas a artigos cuja descrição continha o produto tijolo. De facto, como seria de esperar a maioria dos artigos que mencionem o produto tijolo é classificado como Alvenaria.

| | | | Nome Produto | |
|---|---|---|---|---|
| | | | tijolo | Grand Total |
| Categoria Nível 1 | Categoria Nível 2 | Categoria Nível 3 | Artigo Fact Count | Artigo Fact Count |
| Produtos e Materiais de Construção | Alvenarias | Blocos | 1 | 1 |
| | | Tijolo | 36 | 36 |
| | | Total | 37 | 37 |
| | Betões, Argamassas e Aditivos | Betões, Argamassas e Aditivos | 1 | 1 |
| | | Total | 1 | 1 |
| | Revestimentos, Isolamentos e Impermeabilizações | Revestimentos Cerâmicos, Porcelanicos e de Barro Cozido | 1 | 1 |
| | | Total | 1 | 1 |
| | Total | | 39 | 39 |
| Sem Categoria | Sem Categoria | Sem Categoria | 26 | 26 |
| | | Total | 26 | 26 |
| | Total | | 26 | 26 |
| Serviços e Subempreitadas | Serralharia, Carpintarias e Cantarias | Pedras Naturais | 1 | 1 |
| | | Total | 1 | 1 |
| | Total | | 1 | 1 |
| Grand Total | | | 66 | 66 |

Figura 3.28 – Categorias atribuídas ao produto tijolo.

A Figura 3.29 mostra as unidades de medida que são utilizadas para descrever dois tipos diferentes de produtos: cimento e ferro. Isto permite concluir que tipos de medidas são mais utilizadas em cada produto específico. No caso do ferro, o mais usual é referir uma determinada quantidade de ferro com base no peso (Kilograma).

| | Nome Produto | | |
|---|---|---|---|
| | cimento | ferro | Grand Total |
| Elem Desc | Artigo Fact Count | Artigo Fact Count | Artigo Fact Count |
| Kilograma | 2 | 22 | 24 |
| Metro Linear | | 5 | 5 |
| Palete | 2 | | 2 |
| Saco | 3 | | 3 |
| Tonelada | | 1 | 1 |
| Unidades | 14 | 18 | 32 |
| Grand Total | 21 | 46 | 67 |

Figura 3.29 - Unidades de medida utilizadas em dois produtos diferentes.

## 3.8 Problemas de Qualidade dos Dados na Extracção de Informação do Texto

*Data Quality* (qualidade dos dados) pode ser visto como um conjunto de processos que visam garantir que os dados armazenados sejam: correctos; precisos; consistentes; completos; integrados; adequados às regras de negócio; e adequados aos domínios [33]. Muitas vezes é necessário tratar os dados que estão inconsistentes, sem nunca perder a informação original que é relevante e que melhor representa o universo dos textos estudados.

Relativamente aos textos que foram analisados no âmbito desta tese, os passos 1 e 2 da Figura 3.20 foram os momentos principais onde tiveram de ser tomadas opções, relativamente aos dados processados e filtrados. Estes momentos são relativos ao passo 1 e 2, representado na Figura 3.20. O passo 1 corresponde ao processo de extrair a informação contida na base de dados. Neste caso houve um levantamento dos dados relevantes para o estudo desta tese, nomeadamente os campos das tabelas onde estão descritos os artigos, por parte dos utilizadores. Este processo não foi directo, visto que o modelo de dados relacional não está organizado segundo os artigos, estando a descrição destes algo dispersa e com dados irrelevantes associados. Assim, este processo teve de filtrar somente os campos que continham o texto relativo à descrição do artigo, incluindo a informação a nível do id do artigo. No passo 2 é feita uma normalização dos dados, onde se elimina a informação, associada à descrição do artigo, que é irrelevante para a sua análise. Neste processo é também feita a segmentação das frases (ver secção 3.1),

onde é necessário que se mantenha a consistência original dos dados, utilizando para isso identificadores associados a cada segmento (que corresponde a uma palavra) da descrição do artigo que foi segmentado.

Houve algumas opções tomadas, relativamente a algumas descrições de artigos. Ocorreram casos de descrições que na realidade serviam simplesmente para descrever capítulos dos artigos, conforme mostra a Tabela 3.30. Nestes casos, e tendo em conta outros campos da tabela que estavam relacionados, foram ignoradas estas linhas. As restantes, que realmente descreviam o artigo foram unidas de forma a formar uma descrição completa do artigo (Tabela 3.1, da secção 3.1).

| id | Descricao_artigo |
|---|---|
| 1 | Cap. 1.1 |
| 2 | Fornecimento de tijolo |
| 3 | com medidas 20x20x30 cm. |
| 4 | Cap. 1.2 |
| 5 | Fornecimento de tijolo |
| 6 | com medidas 35x35x40 cm. |

Tabela 3.30 – Exemplo de parte da tabela que contem as descrições dos artigos.

Desta forma, tiveram de ser utilizadas técnicas que permitiram extrair a informação adequadamente, para que fosse possível organizá-la da melhor forma, de modo a ser posteriormente enquadrada na base de conhecimento da Vortal. O capitulo 4 estuda qual o nível de ruído aliado a esses processos e a sua adequação à informação presente nessa base de conhecimento.

# 4. Trabalho experimental

## 4.1 Plano das experiências

Nas subsecções seguintes são apresentados os resultados das experiências que foram efectuadas. Estes ensaios experimentais foram realizados tendo em conta a análise a três factores principais: a ambiguidade na classificação com etiquetas específicas; a geração de frases a partir das regras de produção das gramáticas; e o reconhecimento (total e parcial) das frases da amostra em estudo, com base nas gramáticas.

Para estes ensaios experimentais foi utilizado um conjunto de textos com 2275 frases no total. Cada uma destas frases representa uma descrição de um determinado artigo ou serviço, conforme explicado em detalhe na secção 3.2. Esta base de conhecimento contém as frases estruturadas no formato TXT/2 e anotadas morfossintácticamente com etiquetas genéricas, utilizando um conjunto de treino do CETEMPublico. Adicionalmente foram automaticamente atribuídas ao texto etiquetas específicas, com base no dicionário apresentado na secção 3.4. Verifica-se que este dicionário, apesar de ter sido criado manualmente, apresenta ambiguidades na classificação (com etiquetas específicas) de algumas palavras, sendo este tema analisado na subsecção 4.2.

***Conjuntos de frases***

Para estes ensaios foram utilizados conjuntos de frases, no formato TXT/2, com 100, 500, 1000 e 2275 frases. Desta forma pretende-se estimar a progressão de algumas estatísticas, com base no tamanho da amostra. Cada um dos conjuntos foi construído copiando as primeiras frases do conjunto total de 2275 frases.

Ou seja, todas as amostras menores são subconjuntos das maiores:

Amostra de 100 frases ⊆ Amostra de 500 frases ⊆ Amostra de 1000 frases ⊆ Amostra de 2275 frases

De notar que o conjunto de 100 frases tem 2.985 palavras; o conjunto de 500 frases tem 9.473 palavras; o conjunto de 1000 frases tem 17.580 palavras; e o conjunto de 2275 frases tem 41.705 palavras.

***Etiquetas genéricas e específicas***

Visto que estas frases estão anotadas com etiquetas genéricas e específicas é possível estimar o grau de ambiguidade das etiquetas específicas. Note-se que por apenas ter sido aplicado um processo de desambiguação às etiquetas genéricas, as palavras ainda são ambíguas relativamente às etiquetas específicas. As etiquetes específicas utilizadas nestes ensaios são as apresentadas na secção 3.4.

Para estimar o grau de ambiguidade na anotação com estas etiquetas é feita uma comparação com as principais etiquetas genéricas: N (Nome), ADJ (Adjectivo), PROP (Nome próprio).

***Gramáticas e regras de produção***

De forma a poder validar a estrutura sintáctica e gerar conhecimento a partir dos textos anotados, foram criadas duas gramáticas, uma com regras de produção baseadas nas etiquetas específicas e outra com regras de produção baseadas nas etiquetas genéricas. Assim, são analisados o número de frases validadas pelas gramáticas e geradas pelas regras de produção, tendo em conta a qualidade das frases geradas automaticamente.

Como a abordagem seguida na geração e reconhecimento de frases tem um carácter específico à base de conhecimento utilizada, as regras de produção das duas gramáticas tentam representar algumas frases tipo. É com base nisso que foram igualmente criadas as etiquetas específicas, de modo a ser possível criar a estrutura das regras de produção que são apresentadas de seguida.

A estrutura das regras de produção da gramática que utiliza as etiquetas específicas pode ser consultada na secção 3.5.

A estrutura das regras de produção da gramática que utiliza as etiquetas genéricas, e que é equivalente à gramática específica, é a seguinte:

| | |
|---|---|
| Regra 9 | N PROP |
| Regra 10 | N de N do tipo ADJ |
| Regra 11 | N de N do tipo ADJ com dimensões PROP |
| Regra 12 | N do tipo ADJ com espessura de PROP e acabamento ADJ |
| Regra 13 | N de N ADJ |
| Regra 14 | N ADJ com PROP de espessura N ADJ |
| Regra 15 | N de N de N |
| Regra 16 | N e N de N |

Tabela 4.1 – Regras de produção da gramática genérica.

Nas secções seguintes são explicados e apresentados em detalhe os ensaios experimentais realizados.

## 4.2 Ambiguidade na classificação com etiquetas específicas

***Objectivos do ensaio***

O objectivo deste ensaio é analisar as palavras que criam ambiguidade aquando da classificação com as etiquetas específicas e analisar a relação entre os dois tipos de etiquetas existentes, específicas e genéricas.

Tendo em conta as etiquetas genéricas, verificou-se que existem palavras que tanto podem corresponder a um produto (PROD), um constituinte de um produto (CO_PROD) ou ainda caracterizar um produto (CA_PROD). Desta forma pode tornar-se ambíguo a classificação de uma determinada palavra.

***Definição da experiência***

Foi efectuada uma contagem ao número de palavras que apresentam as etiquetas genéricas Nome (N) e Adjectivo (ADJ) (atribuídas pelo classificador consoante o

contexto da frase). Desta forma, e considerando uma aproximação da relação entre Produto/Nome (PROD/N,) Constitui_Produto/Adjectivo (CO_PROD/ADJ) e Caracteriza_Produto/Adjectivo (CA_PROD/ADJ), seria possível obter um processo de desambiguação para as etiquetas específicas.

Com base nestes pressupostos foram analisadas as palavras classificadas com Nome (N) e Adjectivo (ADJ) em contextos diferentes. Assim é possível avaliar a frequência destes casos e a sua relevância para auxiliar a desambiguação das etiquetas específicas (Experiencia sobre a ambiguidade das etiquetas específicas).

De seguida são mostrados exemplos de frases onde ocorrem palavras que podem criar ambiguidade na classificação com base em etiquetas específicas.

Exemplos:

Fornecimento e colocação de chapins em mármore **vidraça**. (N e N de N em N **ADJ**)
Aplicação de mosaico **plástico**. (N de N **ADJ**)
Fornecimento de paredes de **plástico** do tipo pladur. (N de N de **N** do tipo PROP)
Fornecimento e colocação de paredes de **plástico** do tipo pladur. (N e N de N de **N** do tipo PROP)
Aplicação de juntas de **plástico** em janelas. (N de N de **N** em N)
Fornecimento de **vidraça**. (N de **N**)

Como podemos ver nos exemplos apresentados, as palavras sublinhadas em certos casos podem ser produtos (PROD) e noutros podem caracterizar um produto (CA_PROD). Desta forma, por exemplo, a palavra plástico deveria ser etiquetada como PROD e CA_PROD onde está classificada, respectivamente com, N e ADJ.

Outro teste realizado neste âmbito foi a contagem das etiquetas genéricas que caracterizam uma palavra anotada com determinada etiqueta específica (Experiência sobre a relação entre as etiquetas genéricas e específicas). Ou seja, tendo em conta uma etiqueta específica, foi analisado o número de ocorrências de cada uma das etiquetas genéricas (Tabela 4.2). Desta forma sabe-se qual (ou quais) a etiqueta genérica que ocorre mais vezes quando se anota uma palavra com uma determinada etiqueta específica. Na

Figura 4.3, para cada etiqueta genérica podemos ver a percentagem das etiquetas específicas que são atribuídas.

Na Figura 4.4 podemos ver os mesmos dados de forma inversa.

***Experiência sobre a ambiguidade das etiquetas específicas***

Tendo em conta a amostra de 1000 frases (número total de palavras: 17.580).

Depois de analisadas as frases sabe-se que 6.690 palavras foram anotadas com etiquetas específicas (das quais 588 são palavras distintas) e 14.162 palavras foram anotadas com as etiquetas genéricas estudadas (das quais 2.162 são palavras distintas).

Na Tabela 4.2 estão listados todos os casos encontrados de palavras distintas classificadas com N e ADJ, e cuja etiqueta específica é sempre a mesma.

| **Palavras** | **Total palavras** | **Classificadas como N** | **Classificadas como ADJ** | **Etiqueta específica atribuída** |
|---|---|---|---|---|
| cortes | 54 | 24 | 30 | SERV |
| fechadura | 6 | 4 | 2 | PROD |
| mosaico[11] | 13 | 11 | 2 | PROD |
| paredes | 38 | 37 | 1 | PROD |
| plástica | 21 | 4 | 17 | CA_PROD |
| plástico | 4 | 3 | 1 | PROD |
| porta | 14 | 13 | 1 | PROD |
| quente | 6 | 3 | 3 | CA_PROD |
| remates | 52 | 8 | 44 | SERV |
| tijolo | 25 | 24 | 1 | PROD |
| tinta | 65 | 52 | 13 | PROD |
| tomadas | 8 | 3 | 5 | PROD |
| vidraça[12] | 10 | 9 | 1 | CA_PROD |
| Total (13 palavras): | 316 | 195 | 121 | |

Tabela 4.2 – Estimativa sobre as possíveis palavras ambíguas.

***Análise dos resultados sobre a ambiguidade das etiquetas específicas***

Tendo em conta os resultados da Tabela 4.2, podemos calcular as seguintes estimativas:

[11] Mosaico: pavimento feito de ladrilhos variegados; embutido de pequenas pedras ou de outras peças, formando determinado desenho.

[12] Vidraça: vidro reduzido a forma laminar.

Total de possíveis ambiguidades,

- considerado o total de palavras com etiquetas específicas: 316 / 6.690 = 5%

- considerando o total de palavras distintas com etiquetas específicas: 13 / 588 = 2%

- considerando o total de palavras do dicionário específico, ou seja,

    Nº pal. distintas com possível ambiguidade / Nº pal. do dicionário específico

    13 / 194 = 7%

Tendo em conta o dicionário específico que foi criado, a taxa de possíveis classificações ambíguas é cerca de 7%. Isto porque existem entradas no dicionário onde uma palavra pode ter duas classificações específicas atribuídas. Estes são os casos onde há duas opções de classificação, contribuindo para casos ambíguos. Assim, este valor estimado, foi calculado pelo número total de palavras, desse dicionário, que apresentavam duas classificações possíveis dividindo pelo número total da ocorrência de todas as palavras do dicionário. Importa referir que as etiquetas das palavras que têm duas classificações específicas possíveis, foram obtidas com base no número de frases em que essas palavras ocorriam com uma determinada categoria associada, e noutros casos foi criada a ambiguidade de forma manual, visto ter havido casos de palavras que justificavam etiquetas específicas adicionais.

Relação entre as etiquetas genéricas,

- considerando a etiqueta ADJ: 121 / 316 = 38%

- considerando a etiqueta N: 195 / 316 = 62%

Isto significa que apesar de a percentagem ser pequena, pode de facto ocorrer casos em que se gera ambiguidade na atribuição de uma etiqueta específica.

***Experiência sobre a relação entre as etiquetas genéricas e específicas***

Tendo em conta a amostra de 2275 frases (número total de palavras: 35.942).

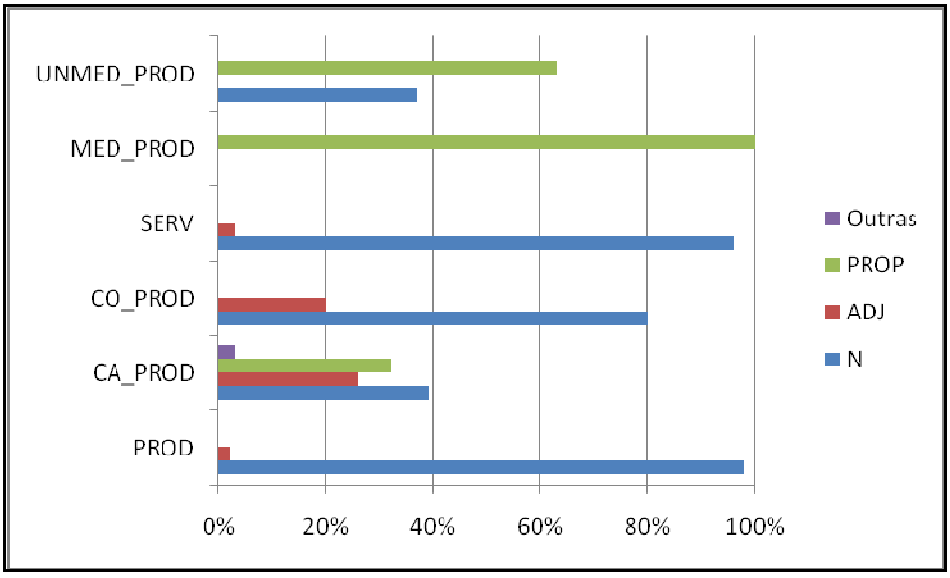



Figura 4.3 – Estimativa da relação entre etiquetas (análise às etiquetas específicas).

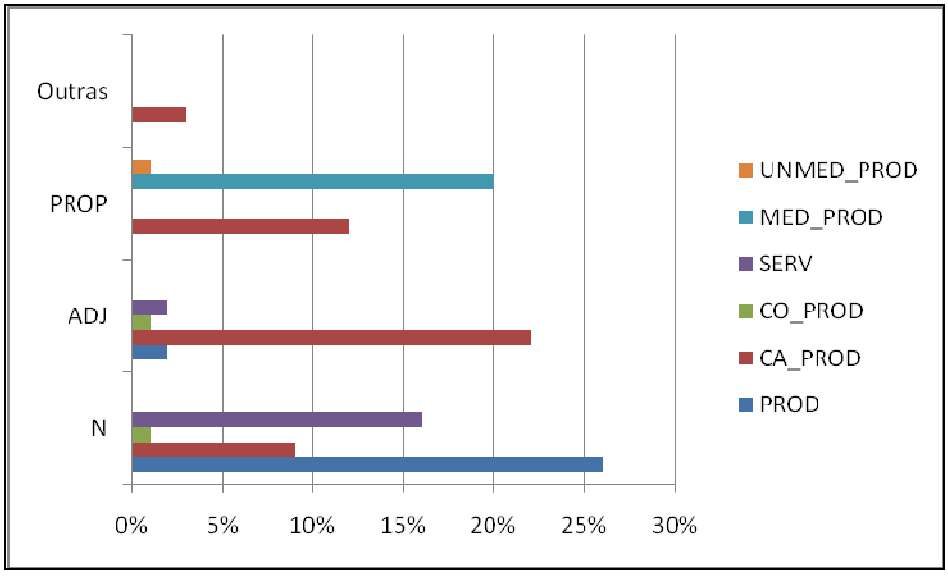



Figura 4.4 – Estimativa da relação entre etiquetas (análise às etiquetas genéricas).

***Análise dos resultados sobre a relação entre as etiquetas genéricas e específicas***

Com base na Figura 4.3, podemos verificar que em todas as palavras anotadas com a etiqueta específica Produto (PROD), 98% foram anotadas com a etiqueta Nome (N) pelo classificador. A etiqueta Serviço (SERV) apresenta resultados similares à etiqueta Produto (PROD), tendo uma correspondência quase total com a etiqueta genérica Nome (N). Outra etiqueta com correspondência total é a Medida do Produto (MED_PROD), ou seja, todas as palavras anotadas com Medida do Produto (MED_PROD) foram anotadas pelo classificador com a etiqueta genérica Nome Próprio (PROP). Nas restantes etiquetas específicas nota-se uma maior dispersão. Desta forma podemos afirmar que na grande generalidade dos casos, todos os Produtos (PROD) são Nome (N) (98%), todos os Constitui Produto (CO_PROD) são Nome (N) (80%) e todos os Serviços (SERV) são Nome (N) (96%).

A Figura 4.4 dá-nos outra visão, do lado das etiquetas genéricas. Assim, de todas as palavras anotadas como sendo Nomes (N), 26% foram anotadas de forma específica como sendo Produto (PROD), 16% como sendo Serviço (SERV) e 47% sem anotação específica. Ou seja, uma palavra que seja um Nome (N) pode ser um Produto (PROD), um Serviço (SERV) ou algo, eventualmente, não contemplado. No caso das palavras anotadas com Adjectivo (ADJ) e com uma etiqueta específica, a quase totalidade das palavras são Caracteriza Produto (CA_PROD).

Analisando agora os resultados das duas tabelas, e tendo como exemplo a anotação com uma etiqueta específica de uma determinada palavra. Se a palavra estiver no dicionário como sendo um Produto (PROD) e um caracteriza Produto (CA_PROD), surge uma ambiguidade. Neste caso podemos tentar desambiguar com base na etiqueta genérica, pois caso esta fosse um Adjectivo (ADJ), é muito mais provável que a palavra fosse na realidade um Caracteriza Produto (CA_PROD) (palavra que caracteriza o produto).

Estes resultados confirmam igualmente que a ambiguidade das gramáticas pode vir a ser reduzida a nível lexical.

*Análise dos resultados sobre a ambiguidade e relação das etiquetas*

Com base na análise dos resultados sobre a ambiguidade e relação entre as etiquetas genéricas e específicas, foi implementado um processo de correcção da anotação das palavras com etiquetas específicas. Isto porque, como vimos anteriormente, existem palavras cuja etiqueta específica não está correcta, tendo em conta o contexto da frase onde está inserida a palavra.

Este processo consiste na desambiguação por regras, onde é tido em conta a classificação feita pelo tagger (aTag) e a classificação feita pelo dicionário específico (userTag). Deste modo, e tendo o auxílio dos Gráficos 4.3 e 4.4 foram criadas as seguintes regras:

A. Se tivermos PROD classificado como N, então é um PROD. (regra sem efeito prático, serve para reforçar o conceito das próximas regras)
B. Se tivermos PROD classificado como ADJ, então é um CA_PROD.
C. Se tivermos SERV classificado como ADJ, então é um CO_PROD.

Depois de implementado este processo foi feita a correcção para a amostra de 2275 frases, originando os seguintes resultados:

- Palavras cuja userTag foi corrigida: 132 palavras

- Palavras cuja userTag foi corrigida para CA_PROD: 57 palavras

- Palavras cuja userTag foi corrigida para CO_PROD: 75 palavras

Exemplo de frase que foi alvo de desambiguação:

**Aplicação de mosaico plástico. (N de N ADJ)**

```
txt([60], [
   w([1, wd='aplicação', lex=[['N',0.9]], atag(rnd)=[['N',0.9]], tag='SERV']),
   w([2, wd='de', lex=[['PRP',0.9]], atag(rnd)=[['PRP',0.9]], tag='PRP']),
   w([3, wd='mosaico', lex=[['N',0.9]], atag(rnd)=[['N',0.9]], tag='PROD']),
   w([4, wd='plástico', lex=[['N',0.5]], atag(rnd)=[['ADJ',0.7]], tag='PROD']),
 []]).
```

## 4.3 Análise à geração de frases

### *Objectivos do ensaio*

O objectivo principal deste ensaio é analisar o número de resultados e a sua qualidade, na geração de frases. Esta análise é feita para o caso das etiquetas específicas e para o caso das etiquetas genéricas, medindo quantitativamente de que forma a utilização de cada um destes tipos de etiquetas influencia no número de análises e geração de possíveis frases.

A análise, ao ser feita para diferentes amostras de textos (100 frases, 500 frases e 1000 frases), permite verificar a progressão da qualidade das frases geradas, tendo em conta o aumento de informação disponível.

Na geração de frases são utilizadas algumas das regras de produção referidas na secção 4.1. Uma parte destas regras utilizam etiquetas específicas (secção 3.5) e a outra parte etiquetas genéricas (secção 4.1), para que se possam tirar conclusões relativamente á utilização destes dois tipos de regras.

### *Definição da experiência*

Numa primeira fase foi necessário proceder à contagem de palavras distintas da amostra, anotadas com cada uma das etiquetas específicas e das etiquetas genéricas, utilizadas nas regras de produção (Tabela 4.5 e Tabela 4.6).

Uma vez que as regras de produção geram as frases com base na anotação feita às palavras, o número de resultados está directamente relacionado com o número de etiquetas para cada regra.

Tendo cada regra de produção várias variáveis relacionadas com o número de palavras distintas anotadas com determinada etiqueta, a produção total destas regras será a combinação entre o produto de cada uma destas variáveis, como é exemplificado de seguida para uma frase que tenha a regra de produção "**Produto Caracteriza_Produto**":

| Produto | Caracteriza Produto |
|---|---|
| Mosaico | plástico |
| Tijolo | vidrado |
| Cimento | metálico |

Tendo duas variáveis que produzem três resultados cada uma, o resultado final gerado será:

*Produto*, *Caracteriza Produto* = 3 X 3 = 9 frases diferentes

Mosaico plástico; Mosaico vidrado; Mosaico metálico; Tijolo plástico; …

Esta é a forma de calcular o número total de resultados que obtemos para cada regra de produção. Não sendo à partida o resultado óptimo, importa inferir sobre a qualidade do mesmo. Para isso é necessário determinar quais destas frases geradas são válidas e inválidas, sendo que uma frase inválida é uma frase que está sintacticamente correcta, mas semanticamente não está correcta, porque não faz sentido.

Visto que na maioria dos casos os resultados gerados eram demasiados para se poder determinar o número de frases válidas analisando individualmente cada uma delas, foi considerada uma ponderação que consistiu em determinar as palavras válidas em cada nível da regra de produção, e no fim calcular o número total das combinações possíveis. Ou seja, no caso da regra de produção: *Produto*, *Caracteriza_Produto*; são determinadas as palavras mais prováveis de gerar uma frase válida para a regra *Produto* e depois para a regra *Caracteriza_Produto*. Finalmente calcula-se as combinações possíveis para as frases geradas, com base no número de palavras escolhidas como sendo válidas para formar uma frase com sentido.

***Resultados experimentais à geração de frases***

| **Amostra (X frases)** | **PROD** | **CA_PROD** | **CO_PROD** | **MED_PROD** | **UNMED_PROD** | **PRP** | **SERV** | ***Total*** |
|---|---|---|---|---|---|---|---|---|
| 100 | 26 | 24 | 3 | 42 | 2 | 5 | 14 | 116 |
| 500 | 72 | 56 | 7 | 209 | 4 | 6 | 29 | 383 |
| 1000 | 89 | 83 | 10 | 355 | 7 | 6 | 38 | 588 |
| 2275 | 97 | 100 | 14 | 590 | 8 | 7 | 40 | 856 |

Tabela 4.5 – Palavras distintas da amostra, anotadas com etiquetas específicas.

| Amostra (X frases) | N | ADJ | PRP | PROP | *Total* |
|---|---|---|---|---|---|
| 100 | 124 | 47 | 7 | 146 | 324 |
| 500 | 445 | 164 | 12 | 732 | 1.353 |
| 1000 | 675 | 273 | 15 | 1.199 | 2.162 |
| 2275 | 1.143 | 513 | 20 | 1.993 | 3.669 |

Tabela 4.6 – Palavras distintas da amostra, anotadas com etiquetas genéricas.

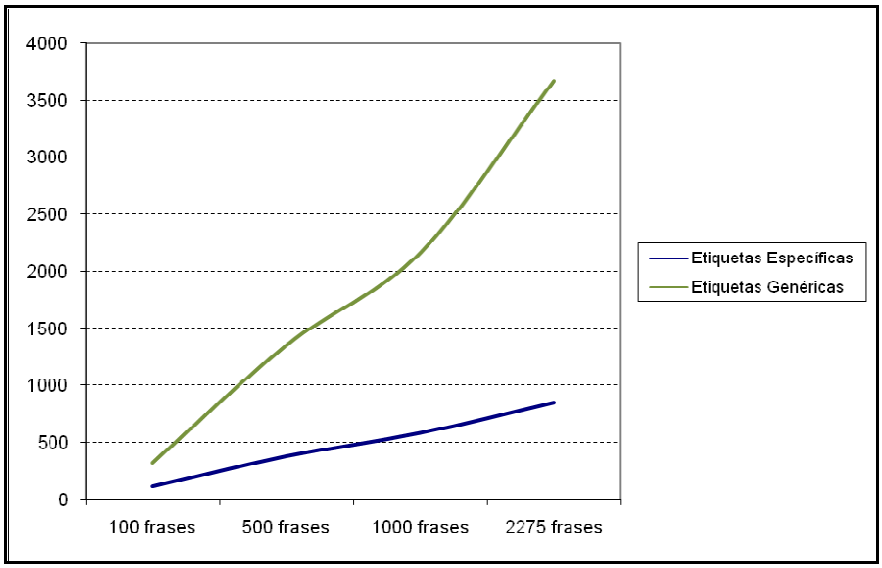



Figura 4.7 – Variação do nº de palavras distintas, anotadas com etiquetas genéricas e específicas.

| | Amostra com 100 frases | | Amostra com 500 frases | | Amostra com 1000 frases | |
|---|---|---|---|---|---|---|
| **Regras de Produção** | **Frases possíveis** | **Frases válidas** | **Frases possíveis** | **Frases válidas** | **Frases possíveis** | **Frases válidas** |
| 1 (et. esp.[13]) | 1.092 | 630 (58%) | 15.048 | 4.140 (30%) | 31.595 | 9.860 (31%) |
| 2 (et. esp.) | 1.872 | 1.688 (90%) | 28.224 | 8.331 (30%) | 73.870 | 26.767 (36%) |
| 3 (et. esp.) | 366.912 | 241.500 (65%) | 24.437.952 | 6.083.040 (25%) | 99.650.630 | 35.290.723 (35%) |
| 4 (et. esp.) | 628.992 | 291.200 (46%) | 47.190.528 | 20.291.927 (43%) | 217.657.955 | 113.182.136 (52%) |
| 5 (et. esp.) | 8.736 | 4.900 (56%) | 116.928 | 59.633 (51%) | 280.706 | 168.423 (60%) |
| 6 (et. esp.) | 8.805.888 | 709.800 (8%) | 1.368.525.312 | 82.111.519 (6%) | 8.271.002.290 | 578.970.160 (7%) |
| 7 (et. esp.) | 1.092 | 672 (62%) | 14.616 | 8.623 (59%) | 33.820 | 22.321 (66%) |
| 8 (et. esp.) | 5.096 | 3.136 (62%) | 60.552 | 34.514 (57%) | 128.516 | 82.250 (64%) |
| 9 (et. gen.[14]) | 18.104 | 2.556 (14%) | 325.740 | 58.633 (18%) | 809.325 | 169.958 (21%) |
| 10 (et. gen.) | 722.672 | 18.900 (3%) | 32.476.100 | 3.261.141 (10%) | 124.385.625 | 7.661.115 (6%) |
| 11 (et. gen.) | 105.510.112 | 954.800 (1%) | 23.772.505.200 | 183.537.900 (1%) | 149.138.364.375 | 1.094.284.620 (1%) |
| 12 (et. gen.) | 39.991.736 | 1.058.400 (3%) | 8.761.103.040 | 87.611.030 (1%) | 60.318.182.925 | 603.181.529 (1%) |
| 13 (et. gen.) | 722.672 | 22.176 (3%) | 32.476.100 | 324.761 (1%) | 124.385.625 | 3.731.569 (3%) |
| 14 (et. gen.) | 4.958.975.264 | 10.378.368 (0%) | 3.898.690.852.800 | 38.986.908.528 (1%) | 40.714.773.474.375 | 407.147.734.744 (1%) |
| 15 (et. gen.) | 1.906.624 | 45.144 (2%) | 88.121.125 | 2.643.634 (3%) | 307.546.875 | 9.226.406 (3%) |
| 16 (et. gen.) | 1.906.624 | 13.376 (1%) | 88.121.125 | 1.762.423 (2%) | 307.546.875 | 9.226.406 (3%) |

Tabela 4.8 – Estimativa da qualidade da geração de frases para cada amostra, com base nas regras de produção.

[13] Etiqueta específica.

[14] Etiqueta genérica.

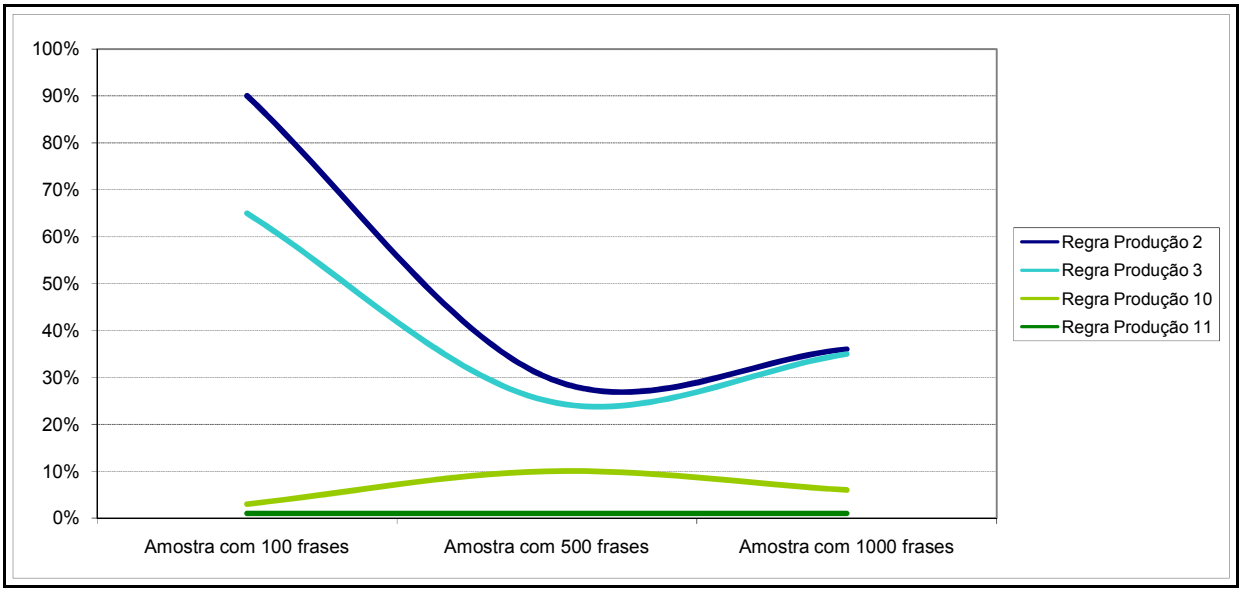



Figura 4.9 – Qualidade dos resultados relacionado com o tamanho da amostra.

***Análise dos resultados sobre a geração de frases***

Relativamente à Tabela 4.5 e Tabela 4.6, verifica-se a grande diferença na quantidade de palavras distintas marcadas com as etiquetas específicas e genéricas. Estas últimas aparecem mais do triplo do que as etiquetas específicas. Isto deve-se ao facto do processo de marcação com as etiquetas específicas não contemplar uma grande diversidade de palavras diferentes. Desta forma são menos as palavras que contêm uma etiqueta específica associada. Tendo em conta estas duas tabelas, verifica-se ainda que o aumento de palavras distintas marcadas com as etiquetas genéricas sobre um aumento muito superior relativamente às palavras distintas marcadas com etiquetas específicas (Figura 4.7), tendo em conta as diferentes amostras, onde varia o número de frases existentes, e consequentemente o número de palavras.

A Figura 4.9 dá-nos uma visão geral da percentagem de frases válidas que conseguimos obter para algumas regras de produção, mediante a amostra a ser utilizada. Estes resultados estão dependentes de vários factores, tais como o número de palavras

existente, o número de palavras anotadas, a qualidade dessas palavras e a própria estrutura das regras de produção.

Verifica-se uma grande diferença na percentagem de frases válidas utilizando a amostra de 100 frases para a amostra de 500 frases, considerando as regras de produção com etiquetas específicas. Esta característica é bem notória na Figura 4.9. Isto deve-se ao facto de na amostra de 100 frases, o número de palavras distintas anotadas com etiquetas específicas é muito inferior ao caso da amostra com 500 frases, deste modo o número de possibilidades para a geração de frases inválidas é muito superior. Algo que reforça esta ideia é a variação que se nota entre a amostra com 500 frases e a amostra com 1000 frases, onde a variação é muito ligeira, tanto na percentagem de frases válidas como no aumento do número de palavras distintas anotadas com etiquetas específicas. Ou seja, da amostra com 100 frases para a amostra com 500 existe um grande aumento de novas palavras que representam Produtos, Constituintes de Produtos, etc.; mas da amostra de 500 frases para 1000 frases não se nota um aumento tão significativo, acabando por não surgirem muitas frases novas. Neste sentido existe uma melhoria na geração de novas frases.

Outro aspecto visível na Tabela 4.8 é a baixa percentagem de frases válidas geradas pelas regras de produção que utilizam as etiquetas genéricas. Comparando com os resultados obtidos utilizando as regras de produção que usam etiquetas específicas, é um valor consideravelmente baixo. Isto deve-se ao facto da geração de conhecimento por parte das regras de produção ser um processo cujos resultados dependem muito dos dados utilizados pelas regras. Visto que as etiquetas genéricas anotam um espectro muito mais amplo de palavras, a probabilidade de anotar palavras que não contribuam para frases válidas é muito grande. À primeira vista seria de esperar uma melhor qualidade na geração, visto que, apesar de as etiquetas serem muito genéricas, os textos anotados têm todos uma linguagem muito específica e relacionada com materiais de construção. Isto advém do facto de não existir uma grande especificidade na relação destas etiquetas genéricas e as regras de produção. Já no caso das etiquetas específicas, estas têm uma grande afinidade com as regras de produção, pois têm uma especificidade propositada para se adequarem às necessidades da geração de frases com base nessas regras, ou seja,

são etiquetas muito específicas para possibilitar uma composição de regras que gerem um número de frases o mais limitado e exclusivas possível.

Ainda na Tabela 4.8, é notória a diferença de frases válidas geradas por cada regra de produção diferente, ou seja, considerando as regras de produção de 1 a 8 (baseadas em etiquetas específicas) observa-se que a 2 e a 3 apresentam uma elevada taxa de frases válidas comparando com as outras regras. Isto deve-se ao facto que quanto mais simples e directa for a regra melhores resultados obtemos na geração de frases com sentido lógico.

A regra 6 apresenta poucas frases válidas devido ao elevado número de etiquetas utilizadas na estrutura da regra, o que faz com que exista um grande número de combinações de frases possíveis, aumentando assim a possibilidade de ocorrerem frases sem sentido. Esta justificação serve igualmente para o caso da regra 14.

## 4.4 Análise ao reconhecimento de frases

***Objectivos do ensaio***

O objectivo deste ensaio é inferir sobre a qualidade das gramáticas e regras de produção, com base no reconhecimento das frases. Para isso é necessário saber qual o número de frases que apresentam uma correcta estrutura sintáctica, tendo em conta as gramáticas definidas e as suas regras de produção.

Visto existirem gramáticas definidas com etiquetas genéricas e com etiquetas específicas, importa comparar a capacidade e facilidade de reconhecimento de ambas as abordagens.

***Definição da experiência***

Para esta experiência foi utilizada a amostra de 2275 frases e o conjunto completo das gramáticas e regras de produção referidas na secção 3.5 e 4.1.

A ideia geral nesta experiência é validar cada uma das frases com base no conjunto completo das gramáticas e suas regras de produção.

Visto que as regras de produção não contemplam os símbolos de pontuação (. , ; : ! ? - « » < >), foram eliminados este tipos de símbolos das frases.

Outro aspecto importante é o tipo de validação que é feita a cada frase, com base nas gramáticas, visto que numa primeira passagem é feito um *matching* entre a frase completa por todas as gramáticas existentes e caso esse *matching* falhe é tentado novamente mas reduzindo a frase a menos uma palavra. Como se pode ver na Figura 4.10, ao reduzirmos a frase surgem várias combinações possíveis para a frase que foi reduzida. Basta haver uma combinação que seja validada pelas regras da gramática para considerar a frase reduzida como sendo válida.

Este processo é repetido até que a frase seja validada por alguma regra da gramática ou quando já não podemos reduzir mais a frase, ou seja, quando reduzimos no mínimo até uma palavra.

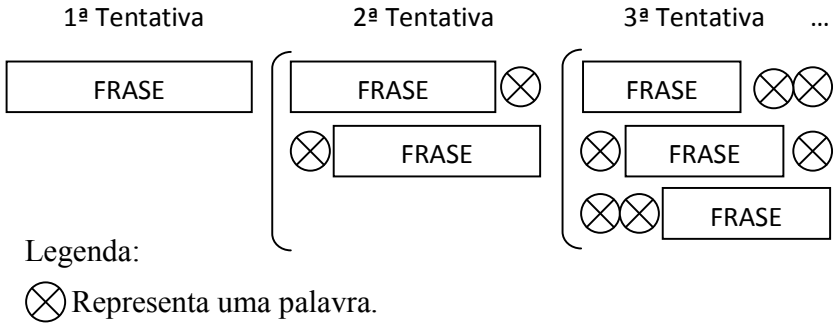



Figura 4.10 – Processo de validação de uma frase.

*Experiência sobre o reconhecimento de frases*

| | | |
|---|---|---|
| **Total de Frases Analisadas:** | **2275** | **frases** |
| **Total Frases Reconhecidas pela gramática genérica:** | **1750** | **frases** |
| **Total Frases Reconhecidas pela gramática específica:** | **438** | **frases** |

| | | |
|---|---|---|
| **% estimada de Frases Reconhecidas:** | **77%** | |
| - Média do Comprimento Reconhecido: | 41% | |

| | | |
|---|---|---|
| **Frases Reconhecidas Parcialmente:** | | |
| - Pelas regras da gramática genérica: | 1664 | frases |
| - Reconhecidas pelo início: | 891 | frases |
| - Reconhecidas pelo meio: | 687 | frases |
| - Reconhecidas pelo fim: | 86 | frases |
| - Pelas regras da gramática específica: | 413 | frases |
| - Reconhecidas pelo início: | 283 | frases |
| - Reconhecidas pelo meio: | 113 | frases |
| - Reconhecidas pelo fim: | 17 | frases |

| | | |
|---|---|---|
| **Frases Reconhecidas Totalmente:** | | |
| - Pelas regras da gramática genérica: | 86 | frases |
| - Pelas regras da gramática específica: | 25 | frases |

| | | |
|---|---|---|
| **Regras que reconheceram as Frases:** | | |
| - artigoProd | 12% | das frases existentes |
| - idG(1) | 4% | |
| - idG(2) | 0% | |
| - idG(3) | 0% | |
| - idG(4) | 0% | |
| - idG(5) | 1% | |
| - idG(6) | 8% | |
| - artigoServ | 7% | das frases existentes |
| - idG(7) | 0% | |
| - idG(8) | 7% | |
| - taggerGeneric | 77% | das frases existentes |
| - idG(9) | 49% | |
| - idG(10) | 0% | |
| - idG(11) | 0% | |
| - idG(12) | 0% | |
| - idG(13) | 7% | |
| - idG(14) | 4% | |
| - idG(15) | 3% | |
| - idG(16) | 15% | |

Tabela 4.11 – Frases reconhecidas pelas regras de produção.

Na Figura 4.12 pode-se observar a relação que existe entre o reconhecimento que é feito pelas regras dos dois tipos de gramáticas criadas: regras produzidas com etiquetas genéricas e regras produzidas com etiquetas específicas.

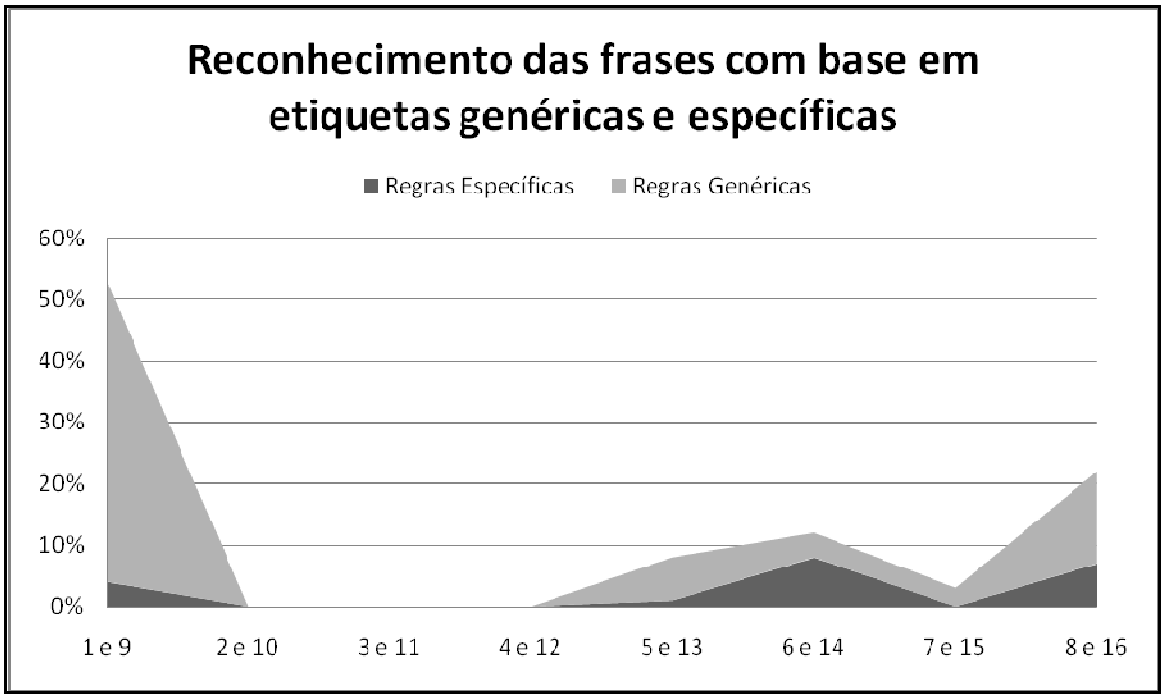



Figura 4.12 – Relação entre as Regras das Gramáticas específicas e genéricas. No eixo das abcissas apresenta-se os identificadores das regras gramaticais.

***Análise dos resultados sobre o reconhecimento de frases***

Verifica-se que 77% das frases da amostra foram reconhecidas por alguma regra da gramática, sendo que em média o comprimento das frases validadas é de 41%. As regras das gramáticas criada com etiquetas específicas validaram 438 frases, estando estas últimas contidas no conjunto de 1750 frases reconhecidas pela gramática criada com etiquetas genéricas.

Isto significa que para reconhecer as frases foi necessário frequentemente diminuir o tamanho das mesmas (removendo as últimas palavras da frase – em média removendo 59%

das palavras em cada frase). Apesar disto, 5% das frases conseguiram ser reconhecidas na sua totalidade.

Existe uma grande diferença no reconhecimento das frases utilizando as gramáticas com etiquetas genéricas face às etiquetas específicas. A taxa de reconhecimento de frases válidas com etiquetas genéricas é superior à com etiquetas específicas. Estes resultados são justificados com base nas etiquetas genéricas que são atribuídas a praticamente todas as palavras das frases existentes, ao contrário do que acontece com as etiquetas específicas, que cobrem um número mais limitado de palavras-chave.

Relativamente à regra que validou mais frases, justifica-se este facto, visto ter características mais genéricas relativamente à estrutura da frase, conseguindo contemplar um maior número de frases ou sub-frases.

Um aspecto que diferencia muito os dois tipos de gramáticas é o possível condicionamento da semântica, por parte da sintaxe, visto que a gramática criada com etiquetas genéricas acaba por ter uma estrutura puramente sintáctica. No caso da gramática com etiquetas específicas, esta já exprime algumas relações semânticas e não apenas sintácticas, visto que as etiquetas têm características relacionais entre si.

Ao ser feita uma avaliação à qualidade semântica das frases reconhecidas pelas regras das duas gramáticas, verifica-se que as regras baseadas em etiquetas genéricas reconhecem uma maior quantidade de frases incorrectas semanticamente, mas correctas a nível de sintaxe.

## 4.5 Análise da qualidade da atribuição das etiquetas específicas

### *Objectivos do ensaio*

O objectivo deste ensaio é determinar a qualidade da atribuição das etiquetas específicas. Esta atribuição, inicialmente, foi feita com base num dicionário de palavras mais frequentes e com um filtro de expressões regulares, como foi explicado na secção 3.4.

Na atribuição de etiquetas específicas às palavras, podem ocorrer situações em que algumas dessas palavras, tendo em conta o seu contexto, podem ter um significado diferente. Como foi discutido na secção 4.2, houve um esforço inicial de desambiguar a primeira atribuição de etiquetas específicas feita às palavras, com base nas etiquetas genéricas (secção 4.2).

Apesar deste processo de desambiguação, pretende-se com este ensaio determinar quantos casos ambíguos ainda existem. Outro aspecto também contemplado, é a existência de palavras que não foram classificadas com nenhuma etiqueta específica, mas que podem ser consideradas como tendo um carácter bastante específico, logo deveriam estar assinaladas como tal.

### *Definição da experiência*

Neste sentido foi efectuada a validação manual das etiquetas específicas a 227 frases da amostra total de 2275 frases. Deste modo, à amostra de 2275 frases (com 41.705 palavras) foram retiradas 227 frases (com 4493 palavras). Utilizou-se o id dos txt/2 para tentar retirar uma amostra não enviesada. Para isso retiraram-se os txt/2 cujo id é múltiplo de 10, obtendo assim uma amostra mais reduzida e dispersa. Como resultado foram obtidos dois conjuntos de frases:

- txt2_fs2275_artigos_to227_CorrigidoMaoComTagInfo.pl
- txt2_fs2275_artigos_to2048_NaoCorrigidoMao.pl

Na correcção manual realizada foram consideradas três condições:

- A etiqueta específica está correcta;
- A etiqueta específica não foi bem atribuída;
- A palavra não tem uma etiqueta específica atribuída, mas é uma palavra relevante, logo é indicada a etiqueta específica correcta.

***Experiência sobre a ambiguidade das etiquetas específicas***

Tendo em conta a amostra de 227 frases, 1780 palavras dessa amostra têm atribuída uma etiqueta específica e 2713 palavras não têm atribuída qualquer etiqueta específica.

| Etiquetas erradas | | |
|---|---|---|
| | | |
| Total de etiquetas erradas: | | 57 palavras |
| | | |
| Etiqueta mal atribuída: | Etiqueta correcta: | Nº de ocorrências: |
| CA_PROD | PROD | 1 palavra |
| CO_PROD | PROD | 1 palavra |
| CO_PROD | SERV | 6 palavras |
| PROD | CA_PROD | 44 palavras |
| SERV | CA_PROD | 5 palavras |

Tabela 4.13 – Etiquetas erradas

| Etiquetas certas | |
|---|---|
| | |
| Total de etiquetas certas: | 1723 palavras |
| | |
| Etiqueta correcta: | Nº de ocorrências: |
| CA_PROD | 332 palavras |
| CO_PROD | 14 palavras |
| MED_PROD | 172 palavras |
| PROD | 313 palavras |
| PRP (ou palavra de ligação) | 602 palavras |
| SERV | 272 palavras |
| UNMED_PROD | 18 palavras |

Tabela 4.14 – Etiquetas certas

| Novas etiquetas atribuídas | | |
|---|---|---|
| | | |
| Total de palavras com novas etiquetas atribuídas: | 21 | palavras |
| | | |
| Etiqueta atribuída: | Nº de ocorrências: | |
| CA_PROD | 14 | palavras |
| PROD | 6 | palavras |
| SERV | 1 | palavra |

Tabela 4.15 – Novas etiquetas atribuídas

***Análise dos resultados sobre a qualidade da atribuição das etiquetas específicas***

Tendo em conta que, da amostra total de 4493 palavras, 1780 são palavras que têm atribuída uma etiqueta específica, a estimativa de erro na atribuição automática destas etiquetas é de cerca de 3%, resultando numa qualidade de 97% na correcta atribuição das etiquetas específicas. É importante ter em conta o processo que foi utilizado para atribuir automaticamente estas etiquetas, processo este, explicado na secção 3.4. De notar que os principais factores para desambiguar utilizados na amostra estudada são as regras impostas com base nas etiquetas genéricas (ver pág. 64, Análise dos resultados sobre a ambiguidade e relação das etiquetas).

Neste estudo é possível verificar que a etiqueta específica Produto (PROD) é a que mais vezes é atribuída incorrectamente (Tabela 4.13). Este facto é explicado com base na Experiência sobre a ambiguidade das etiquetas específicas (secção 4.2), isto porque, considerando os dados da Tabela 4.2 verifica-se que as palavras inicialmente classificadas como Produto (PROD) têm uma maior propensão para serem ambíguas. Isto porque, estas palavras podem ser Nomes (N) ou Adjectivos (ADJ), consoante o contexto, o que faz com que deixem de ser consideradas como Produto (PROD). Uma possível melhoria seria criar mais regras relacionadas com o contexto de cada palavra, utilizando para isso as etiquetas genéricas, da mesma forma que foi explicado na secção 4.2. Outro factor que pode influenciar a qualidade destas etiquetas é a granularidade das mesmas, isto porque, ao termos etiquetas cada vez mais específicas, estamos a diminuir a ambiguidade que as mesmas podem ter.

Na Tabela 4.13 podemos ainda ver para cada etiqueta específica mal atribuída, qual a etiqueta específica que seria a adequada. Esta informação ajuda a perceber que todas as palavras com etiquetas Produto (PROD) mal atribuídas, deviam ter como etiqueta específica Caracteriza Produto (CA_PROD).

Relativamente às palavras cuja atribuição da etiqueta estava certa (Tabela 4.14), verifica-se que ocorre de uma forma homogénea por todas as etiquetas específicas utilizadas.

Durante a validação manual que foi feita às etiquetas específicas, foram também anotadas palavras que não tinham associada qualquer etiqueta específica (Tabela 4.15). Isto porque, algumas palavras foram consideradas relevantes e com um carácter específico e que se ajustavam a algumas das etiquetas específicas criadas. Estas palavras não foram anotadas com as etiquetas específicas, porque não foram contempladas no dicionário inicial que foi criado (secção 3.4).

## 4.6 Considerações finais sobre as experiências

De uma forma geral, e tendo em conta os vários ensaios experimentais efectuados, verifica-se uma grande vantagem na utilização de etiquetas específicas, quando os dados textuais em causa têm uma grande especificidade relativamente a determinados termos e palavras, como é o caso da área da construção. Apesar de mesmo fazendo uma boa classificação das palavras, acabam por existir sempre casos ambíguos, em que consoante o contexto da palavra esta deveria ser classificada com uma etiqueta específica diferente. Nestes casos, o auxílio das etiquetas genéricas pode melhorar a qualidade da anotação com etiquetas específicas.

Os resultados obtidos ainda estão longe de serem os ideais, devendo-se este facto, sobretudo ao dicionário com a listagem de palavras associadas com as etiquetas específicas e às regras de produção das gramáticas. Para melhorar estes aspectos terá de ser feito um levantamento mais exaustivo e rigoroso de regras de produção ainda mais específicas, o que implica igualmente a criação de novas etiquetas específicas, permitindo

a produção de gramáticas com regras mais complexas. Outro aspecto será melhorar as regras de desambiguação, visto se ter obtido alguns resultados positivos com base neste procedimento.

O teste feito ao reconhecimento das frases com base nas gramáticas, mostrou ser uma forma de se poder validar eventuais inserções de novas frases. Desta forma, tendo os mesmos recursos a nível de regras e classificação de textos, pode-se gerar ou reconhecer conhecimento. Este conhecimento corresponde nesta fase inicial à geração de frases com uma estrutura que seja validada pelas regras das gramáticas e utilizando palavras conhecidas e classificadas como sendo da área da construção e/ou de outras áreas específicas.

# 5. Conclusões

## 5.1 Principais resultados

Este estudo teve como principal resultado a ligação entre textos que descrevem artigos, de produtos da construção civil, e uma base de conhecimento no contexto da plataforma Vortal. Para tal, foram apresentadas técnicas de processamento de língua natural contextualizadas neste sistema de informação específico. Esta ligação permitiu analisar por um lado a utilidade em anotar os textos morfossintácticamente estruturando-os com base nas etiquetas, e por outro lado, a adaptação das técnicas de processamento do português a temas específicos, como é o caso da construção civil.

Tendo os textos anotados, tem-se a possibilidade de construir uma base de conhecimento mais estruturada do que a existente actualmente. Isto significa que, em vez de se terem os textos simplesmente guardados, e com um carácter meramente informativo, a anotação traz-nos vantagens relativamente ao aproveitamento da informação que á partida não poderia ser directamente utilizada nos processos de *Business Intelligence*. Permite-se assim, um maior tratamento e manipulação dos dados, passando a possibilitar a obtenção de dados estatísticos sobre os produtos, as referências dos produtos, as medidas e dimensões mais usuais, as características dos produtos, etc. Isto poderá trazer vantagens, a um nível mais geral, para as empresas que controlam a troca de informação, ou a um nível mais específico, para os utilizadores que interagem directamente com os sistemas e com os dados. Para estes utilizadores, é possível auxiliar a escrita de novas descrições de artigos, validando e dando propostas de expressões.

No que respeita ao processamento dos textos e às técnicas utilizadas na anotação, houve a necessidade de detalhar mais as palavras classificadas. Neste sentido foram criadas etiquetas específicas, que demonstraram ser de grande utilidade para a

compreensão da estrutura das frases, e obtenção de bons resultados na geração de frases, com base nas regras gramaticais específicas.

Tendo em conta estes aspectos, surge a possibilidade de relacionar a extracção de informação com sistemas de categorização mais refinados, possibilitando relações entre as anotações dos textos e sistemas de categorias dos produtos e serviços da construção civil. Ainda relacionado com esta questão da análise dos dados, estão as ontologias e a aplicabilidade. Deste modo, foi possível a criação de um *Data Warehouse*, que permite gerar estatísticas sobre toda a informação da base de conhecimento relacionada com os artigos. Esta integração torna-se especialmente relevante quando temos em conta o sistema de categorias (actualmente existente na Vortal) ou sistemas como os referidos na secção 2.1. Estes dois aspectos demonstraram-se particularmente importantes, na medida em que, proporcionam uma estruturação dos dados e uma integração dos mesmos com as ferramentas de análise utilizadas. Inerente a isto está o conceito B2B e especialmente a arquitectura SOA, funcionando esta como ferramenta que ajuda a promover a aplicabilidade e interoperabilidade, demonstrando a utilidade dos módulos actualmente existentes na plataforma Vortal.

## 5.2 Trabalho futuro

Um dos aspectos principais que pode ser alvo de posteriores estudos, tem a ver com o concretizar de modelos que relacionem o conhecimento extraído dos textos com o sistema de classificação de produtos e serviços já existente na Vortal. Desta forma, poder-se-ia estudar a ligação das ontologias com as categorias, numa vertente em que se avaliava a capacidade de ligação entre as frases reconhecidas pelas regras das gramáticas com a categoria, podendo assim inferir sobre a validade da classificação das frases com base no sistema de categorias. Isto porque esta indicação da categoria do produto descrito é feita manualmente, e como tal, podem haver casos menos correctos. Tendo esta relação entre as ontologias e o sistema de categorias, era possível corrigir este tipo de problemas.

Relativamente à relação entre as etiquetas e a informação extraída do texto, gera-se aqui uma ontologia. Isto na medida em que passam a existir várias estruturas de

conhecimento bem organizadas, como é o caso das regras gramaticais, definidas com base nas etiquetas. Deste modo, estas ontologias são passíveis de serem confrontadas com um sistema de categorização de artigos, criando uma ligação mais forte entre os dados que estão na plataforma e a base de conhecimento estruturada. Assim, podem-se relacionar as gramáticas que analisam e validam as frases, com ontologias ou grupos de utilizadores. Deste modo, pode ser estudado um possível processo que permita determinar a que grupo o utilizador pertence, com base no histórico de textos redigidos. Isto pode ser especialmente útil para mostrar ou fornecer conteúdos específicos (publicidade direccionada, etc.) a cada tipo de utilizador. Neste quadro serão particularmente relevantes sistemas como os referidos na secção 2.2.1.

## 5.3 Modelo de uma ontologia para os dados

A Figura exemplo sugerido na secção 3.7, para organizar a informação, tende a ter uma estrutura bem definida a nível de contexto dos dados. Nesta fase é possível desenvolver modelos ontológicos que caracterizem os dados de uma forma relacional e contextual.

Isto não implica que os dados estejam normalizados, visto que as ferramentas OLAP (*Online Analytical Processing*) são geralmente desenvolvidas para trabalhar com bases de dados não normalizadas. Nas ferramentas de análise OLAP, é possível navegar entre diferentes níveis de detalhe (granularidade) de um cubo de dados, semelhante ao apresentado na secção 3.7. Isto através de um processo chamado *Drill*, o utilizador pode aumentar (*Drill down* - detalhar) ou diminuir (*Drill up* - condensar, resumir) o nível de detalhe dos dados.

Este processo é também viável seguindo o modelo de uma ontologia, visto que é possível navegar através dos diversos conceitos da ontologia, e ir aumentando o nível de detalhe de cada conceito (secção 2.1).

Embora estando para além do âmbito desta tese, é interessante pensar numa perspectiva de descoberta de conhecimento, utilizando, p.ex. árvores de decisão. Estas árvores de decisão são uma forma de representar o conhecimento obtido. As árvores de

decisão podem ser construídas automaticamente partindo de conjuntos de dados supervisionados, através da utilização de algoritmos como o C4.5. São representações simples do conhecimento e um meio eficiente de construir classificadores, que determinam as classes, baseando-se nos valores dos atributos num conjunto de dados [21]. São também um poderoso instrumento para o *Data Mining*, na medida em que possibilitam a análise dos próprios dados.

Como foi visto anteriormente (secção 3.7.1), pode ser criada uma base de conhecimento conjugando informação da base de dados operacional com informação extraída do texto.

Assim, é possível interrogar esta base de conhecimento, de modo a obter informações actuais sobre os dados. É aqui que surge a possibilidade de descoberta de informação. Para isso, é necessário obter informação agregada e potencialmente útil para um processo de *Data Mining*. Deste modo, podem ser utilizados modelos baseados utilizando a base de conhecimento da empresa para agrupar e relacionar vários dos elementos detectados no texto (e.x. produtos, características, medidas, locais). Torna-se assim possível pensar, p.ex. na construção automática de árvores de decisão, conjugando informação relacional com a informação representada em fontes textuais. A utilização e análise destes modelos híbridos possibilitará ao analista humano maior controlo e acesso do conhecimento já existente na empresa.

# 6. Bibliografia


[1] AMORIM, S.R.L.; CHERIAF, M.; **Sistema de indexação e recuperação de informação em Construção baseado em Ontologia**. III Encontro Tecnologia da Informação e Comunicação na Construção Civil - TIC2007. Porto Alegre : NORIE UFRGS, 2007.

[2] AMORIM, S.R.L.; PEIXOTO, L.A.; **CDCON: classificação e terminologia para a construção**. Coletânea Habitare – vol. 6 – Inovação Tecnológica na Construção Habitacional, Cap. 8, pág. 188 – 217, Porto Alegre: ATAC, 2006.

[3] AMORIM, S.R.L.; PEIXOTO, L.A.; **Desenvolvimento de terminologia e codificação de materiais e serviços para construção**. In: AMORIM, S.R.L (org); BONIN, L.C. (org). Inovação Tecnológica na Construção Habitacional. 1ª ed. Porto Alegre: ANTAC, 2006.

[4] BIEBERSTEIN, N.; **Service-Oriented Architecture (SOA) Compass**. ISBN-13: 978-0131870024, IBM Press, 2005.

[5] BILL, E.; **Transformation-based error-driven learning and natural language processing: A case study in part-of-speech tagging**. Computational Linguistics, ISSN:0891-2017, MIT Press, 1995.

[6] DECLERCK, T.; KRIEGER, H.; SAGGION, H.; SPIES, M.; **Ontology-Driven Human Language Technology for Semantic-Based Business Intelligence**. Proceedings of ECAI, pág. 841, 2008.

[7] FENSEL, D.; BUSSLER, C.; **The Web Service Modeling Framework WSMF**. Electronic Commerce Research and Applications, Volume 1, Issue 2, Pages 113-137, 2002.

[8] FENSEL, D.; **Ontologies: Silver Bullet for Knowledge Management and Electronic Commerce**. Springer-Verlag, Berlin, 2001.

[9] **GATE - General Architecture for Text Engineering**. http://gate.ac.uk, última vez visitado em Junho de 2008.

[10] GONÇALVES, P.; VIEIRA, R.; RINO, L.; **CorrefSum: Referencial Cohesion Recovery in Extractive Summaries**. Proc International Conf. on Computational Processing of Portuguese - PROPOR, Aveiro, Portugal, pp. 224-227, ISBN: 978-3-540-85979-6, Springer, September, 2008.

[11] KAO, A.; POTEET, STEVE, R.; **Natural Language Processing and Text Mining**. Springer, ISBN-10: 184628175X, 2006.

[12] LINARDAKI, E.; **Reducing Bias Effects in DOP Parameter Estimation**. Proceedings of ECAI, pág. 283, 2008.

[13] LOPES, C.; PERDIGÃO, F.; **Event Detection by HMM, SVM and ANN: A Comparative Study**. Proc International Conf. on Computational Processing of Portuguese - PROPOR, Aveiro, Portugal, pp. 1-10, ISBN: 978-3-540-85979-6, Springer, September, 2008.

[14] MARQUES, N.C.; BADER, S.; ROCIO, V.; HOLLDOBLER, S.; **Neuro-Symbolic Word Tagging**. In: José Neves, Manuel Filipe Santos and José Machado (eds), New Trends in Artificial Intelligence, Associação Portuguesa para a Inteligência Artificial (APPIA), Guimarães. Portugal, ISBN-13978-989-9561, December 2007.

[15] MARQUES, N.C.; GONÇALVES, S.; **Applying a Part-of-Speech tagger to Postal Address Detection on the Web**. In Proceedings of the IV International Conference on Language Resources and Evaluation. LREC 2004. Volume I. pp. 287-290. Lisboa, Portugal, ISBN: 2-9517408-1-6, 2004.

[16] MARQUES, N.C.; LOPES, G.P.; **Neural networks, part-of-speech tagging and lexicons**. Number:6 DI - FCT/UNL. 1998. http://linguateca.pt/Repositorio/tr_di6_98.ps

[17] MARQUES, N.C.; LOPES, G.P.; **Tagging With Small Training Corpora**. In Proceedings of the International Conference on Intelligent Data Analysis (IDA'01), pag 63-72. Lecture Notes in Computer Science 2189, Cascais, Portugal. Springer Verlag. Setembro 2001.

[18] MIRROSHANDEL, S.A.; GHASSEM-SANI, G.; **Unsupervised Grammar Induction Using a Parent Based Constituent Context Model**. Proceedings of ECAI, pág. 293, 2008.

[19] MOHAN, C.; **Dynamic e-Business: Trends in Web Services**. In Proceedings of the third VLDB workshop on Technologies for E-Services, number 2444 in LNCS http://www.almaden.ibm.com/u/mohan/WebServices_TES2002.pdf, 2002.

[20] MONTEIRO, A.; BARBAS, J.; MARQUES, N.; **Utilização da programação declarativa para processamento do CETEMPúblico**. In Luís Costa, Diana Santos & Nuno Cardoso (eds.). Perspectivas sobre a Linguateca, Actas do encontro Linguateca: 10 anos. Linguateca. Novembro de 2008. http://www.linguateca.pt/LivroL10/. (ISBN: 978-989-20-1445-6.)

[21] MITCHELL, T.; **Machine Learning**. McGraw-Hill Science/Engineering/Math; 1 edition (March 1, 1997), ISBN-13: 978-0070428072, 1997.

[22] NASCIMENTO, L. A.; **Proposta de um Sistema de Recuperação de Informação para Extranet de Projeto**. Dissertação de Mestrado, Escola Politécnica da Universidade de São Paulo, USP, SP, 2004.

[23] OLIVEIRA, H.; SANTOS, D.; GOMES, P.; SECO, N.; **PAPEL: A Dictionary-Based Lexical Ontology for Portuguese**. Proc International Conf. on Computational Processing of Portuguese - PROPOR, Aveiro, Portugal, pp. 31-40, ISBN: 978-3-540-85979-6, Springer, September, 2008.

[24] PEREIRA, F.C.N.; SHIEBER, S. M.; **Prolog and Natural-Language Analysis**. Digital edition, Microtome Publishing, 2002.

[25] **Projecto Linguateca**. http://www.linguateca.pt, última vez visitado em Julho de 2008.

[26] RABELO, P.F.R.; AMORIM, S.R.L.; LYRIO FILHO, A.M.; **Ontology, management of Project Process and Information Technologies**. In: ECPPM-2006: e-Business and e-Work in Architecture, Engineering and Construction. Valencia, Espanha, 2006.

[27] RABELO, P.F.R.; AMORIM, S.R.L.; **ONTOARQ – Ontologia para Arquitectura, Engenharia e Construção. Visualização e gerenciamento na Web**. III Encontro Tecnologia da Informação e Comunicação na Construção Civil - TIC2007. Porto Alegre : NORIE UFRGS, 2007.

[28] SANTOS, D.; ROCHA, P.; **CETEMpublico**. Linguistic Data Consortium, Philadelphia, http://ldc.upenn.edu/Catalog/CatalogEntry.jsp?catalogId=LDC2001T62, 2001.

[29] SANTOS, E.T.; NASCIMENTO, L.A.; Re**cuperação de Informação em Sistemas de informações na Construção Civil: O caso das extranets de projeto**. Seminário de Tecnologia de Informação e Comunicação na Construção Civil, 2002.

[30] SILVA, João Ricardo M.F.; **Shallow Processing of Portuguese: From Sentence Chunking to Nominal Lemmatization**. Master Thesis, Universidade de Lisboa, 2007.

[31] TENENBAUM, J.; KHARE, R.; **Business Services Networks: Delivering the Promises of B2B**. ACM International Conference Proceeding Series; Vol. 87, Proceedings of the IEEE EEE05 international workshop on Business services networks, 2005.

[32] TREBIEN, E.; ROSSONI, P.; JACOSKI, C.A.; COSTELLA, M.; **Elaboração de um Centro Regional de Informações da Construção – Projeto CinC**. III Encontro Tecnologia da Informação e Comunicação na Construção Civil - TIC2007. Porto Alegre : NORIE UFRGS, 2007.

[33] WANG, R., KON, H.; MADNICK, S.; **Data Quality Requirements Analysis and Modelling**. Ninth International Conference of Data Engineering, Vienna, Austria, 1993. http://web.mit.edu/tdqm/www/tdqmpub/IEEEDEApr93.pdf